RICE UNIVERSITY

Estimating the Term Structure With a Semi-Parametric Bayesian Population Model: An Application to Corporate Bonds

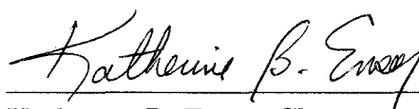
by

Alejandro Cruz Marcelo

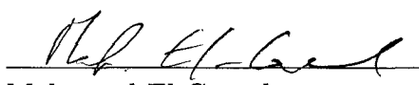
A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

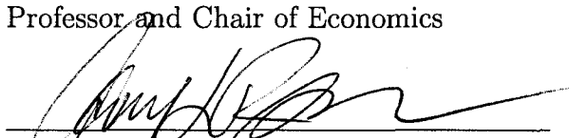
Approved, Thesis Committee:



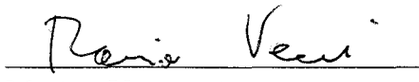
Katherine B. Ensor, Chair
Professor and Chair of Statistics



Mahmoud El-Gamal
Professor and Chair of Economics



Gary L. Rosner
Adjunct Professor of Statistics



Marina Vannucci
Professor of Statistics

Houston, Texas

April, 2010

ABSTRACT

Estimating the Term Structure With a Semi-Parametric Bayesian Population Model: An Application to Corporate Bonds

by

Alejandro Cruz Marcelo

The term structure of interest rates is used to price defaultable bonds and credit derivatives, as well as to infer the quality of bonds for risk management purposes. We introduce a new framework for estimating the term structure of interest rates for corporate bonds. The proposed model jointly estimates term structures by means of a Bayesian hierarchical model with a non-parametric prior probability model based on Dirichlet process mixtures. The main advantage of our framework is its ability to produce reliable estimators at the company level even when there are only a few bonds per company. The modeling methodology borrows strength across similar term structures for purposes of estimation. After describing the new approach, we discuss an empirical application in which the term structure of 197 individual companies is estimated. The sample of 197 consists of 143 companies with only one or two bonds. In-sample and out-of-sample tests indicate superior performance of our method as compared with the popular approach of grouping the corporate bonds by credit rating.

We also discuss the relative performance of different modeling strategies that introduce dependence on covariates into Bayesian nonparametric models. We show that 1) nonparametric models using different strategies for modeling covariates can show noteworthy differences when they are being used for prediction, even though

they produce otherwise similar posterior inference results, and 2) when the predictive density is a mixture, it is convenient to make the weights depend on the covariates in order to produce better estimators. Such claims are supported by comparing the Linear DDP (an extension of the Sethuraman representation) and the Conditional DP (which augments the nonparametric distribution to include the covariates); we apply those methods to a simulated data set and to data from a pharmacokinetic meta-analysis.

Acknowledgments

The completion of this thesis would not have been possible without the guidance and support of several individuals to whom I would like to extend my sincere appreciation and gratitude. I am deeply grateful to my advisor, Dr. Katherine B. Ensor, for her guidance and constant encouragement. Her patience and supervision as well as her integrity and work ethic from which I have learned and grown personally as well as professionally.

I am deeply grateful to Dr. Gary L. Rosner for providing me with the initial exposure to nonparametric Bayesian methods. I would not have been able to surmount some of the obstacles, in the course of this work, without his thoughtful explanations and proving questions. His stimulating criticism and enthusiasm for discussion made the process intellectually rewarding.

I would like to thank two other members of my thesis committee, Dr. Marina Vannucci and Dr. Mahmoud El-Gamal, for taking their valuable time to read this thesis and making helpful comments and suggestions.

I also thank Dr. Peter Mueller for his prompt and concise output. It was a privilege to have access to an expert on Bayesian methods.

I would like to thank all of the faculty, staff and my colleagues in the Statistics Department for their support and encouragement. Special thanks to Tuan Nguyen and David Kahle for numerous helpful discussions.

I want to express my gratitude to Rice University for financially supporting my graduate studies. Specifically, I want to acknowledge the President's Graduate Fellowship, The Brown Foundation Fellowship, The Center for Computational Finance and Economic Systems, and the NSF Vigre Grant DSM-0739420.

On a personal note, I would like to thank my parents, my brother and my sister for their love, support and devotion over the years. Without them, I would not have been the person I am today.

Contents

Illustrations

Tables

Chapter 1

Introduction

The term structure of interest rates, also called the zero-coupon yield curve, refers to the relationship between the interest rate of zero-coupon bonds and their time to maturity. The term structure can be estimated for government or corporate bonds. The term structure of government bonds, also referred to as risk-free term structure, is important because it contains information about macroeconomic conditions and expectations of market agents about the future economic conditions. On the other hand, the term structure of corporate bonds is an essential input for pricing defaultable bonds and credit derivatives (Jarrow and Turnbull 1995; Duffie and Singleton 1999). In addition, it can be used to infer the credit quality of bonds for risk management purposes (Saunders and Allen 2002) as well as to asses risk in derivative products (Hull and White 1995; Duffee 1996).

We introduce a novel framework to estimate the term structure of interest rates. Our model jointly estimates term structures by means of a Bayesian population model with a prior probability model based on Dirichlet process mixtures. To the best of our knowledge, this is the first model for estimation of the term structure that uses a population model approach as well as the first one using non-parametric Bayesian methods. The most important contribution of our model with respect to current esti-

mation methods is its ability to produce accurate estimators based on small samples. This feature is particularly relevant when estimating the term structure of corporate bonds because it allows us to estimate the term structure of individual firms even though individual corporations issue only a handful of bonds. Our model is the first estimation method that is widely applicable to estimate term structures for bonds of individual firms.

A related problem is to predict the term structure based on a given set of bond characteristics or covariates; those estimators can be used, for example, to price new issues of bonds. We propose to address such an estimation problem by extending our term structure model to introduce dependence on covariates. The use of such an extension motivates the research problem we address in the second half of this document.

Although modeling dependence on covariates in nonparametric Bayesian models has been a very active area of research, limited research has examined the relative performance of such methods or improved understanding of which features are suitable in order to produce better results. We consider such a comparison, focusing on predictive inference, and show that different approaches for modeling dependence on covariates can lead to very similar posterior fits and yet produce very different results when used for prediction. In addition, when the predictive density is a mixture, we show that making the weights depend on the covariates plays a major role in determining the quality of the predictions. Such findings are illustrated by comparing the

linear dependent Dirichlet process (De Iorio et al. 2009) to the conditional Dirichlet process (Müller et al. 1996); we apply those methods to a simulated data set and to data from a pharmacokinetic meta-analysis. An area of future research is to use our findings on modeling dependence on covariates to extend our term structure estimation method so that accurate estimators can be obtained based on a given set of covariates.

This document is organized as follows. Chapter 2 introduces our new framework to estimate the term structure, illustrates the performance of the proposed with an empirical application and a simulation example, and, finally, describes how to extend our term structure model to introduce dependence on covariates. Such an extension motivates the research problem studied in Chapter 3 which refers to the analysis and comparison of modeling approaches that introduce dependence on covariates into nonparametric Bayesian models. Discussion and directions of future work appear in Chapter 4. Appendixes A and B include a description of a MCMC algorithm to implement the proposed term structure model. Appendix C includes some comments and results on a less flexible version of our term structure model in which the prices are assumed to be normally distributed. Finally, Appendix D describes the software implementation of the proposed model.

Chapter 2

Estimating the Term Structure

We present a novel framework to estimate the term structure of interest rates. The proposed model jointly estimates multiple term structures by means of a hierarchical model that resembles a Bayesian population model. A joint model allows us to pool information and borrow strength across the term structures. For the implementation of such a population model approach, it is important to determine what term structures are most likely to have similar characteristics, and consequently, should be borrowing strength from each other. In order to cluster similar term structures, we use a prior based on Dirichlet process mixtures. Such a flexible nonparametric prior probability model also allows us to accommodate the heterogeneity in the population of parameters that characterize the term structures. Manifestations of such heterogeneity includes outliers, over-dispersion, and multimodality. To the best of our knowledge, this is the first model for estimation of the term structure that uses a population model approach as well as nonparametric Bayesian methods.

The most important contribution of our model with respect to current estimation methods is its ability to produce accurate estimators based on small samples. This feature is particularly relevant when estimating the term structure of corporate bonds because it allows us to estimate the term structure of individual firms even though

individual corporations issue only a handful of bonds. Using the estimators produced with our model leads to a remarkable improvement with respect to the single-curve approach which is commonly used to approximate the term structure of corporate bonds. The single-curve approach includes two stages: first, the corporate bonds are classified by credit rating level; then, the term structures are independently estimated for each class. Grouping the bonds by credit rating, and not by issuer, is necessary under the single-curve approach because it guarantees that the resulting groups include enough bonds to apply traditional estimation methods. Our model eliminates the need of grouping the bonds by credit rating and allows practitioners to use the more accurate estimators corresponding to individual issuers.

This chapter is organized as follows. Section 2.1 provides a survey of existing estimation methods of the term structure. Our new framework is described in Section 2.2. In Section 2.3 we apply our model to empirical data to illustrate its performance; this section includes a description of the data set as well as in-sample and out-of-sample tests that compare our approach with the single-curve method. Section 2.4 illustrates the performance of the proposed model by using a simulation example. Finally, conclusions and discussion appear in Section 2.6.

2.1 Estimation Methods

Since most of the corporate and government bonds have a positive coupon, their term structures are not observable and they have to be estimated from market prices

using statistical technique. In this document we consider estimation methods that are based on the discounted cash flow (DCF) principle. This section defines the DCF principle and explains how it has been used to estimate term structures. We conclude with a review of estimation methods for corporate bonds.

Before introducing the DCF principle, we discuss equivalent representations of the term structure. One representation is the zero-coupon yield curve, $y(T)$, which describes the relationship between spot rates of zero-coupon bonds and their time to maturity, T. Two other representations of the term structure are the discount curve, $D(T)$, and the forward rate curve, $f(T)$. The representations $y(T)$, $D(T)$ and $f(T)$ are all equivalent since they satisfy the following relationships:

$$D(T) = exp\{-Ty(T)\} = exp\left\{-\int_0^T f(s)ds\right\}. \tag{2.1}$$

For the derivation of these relationships, see, for example, Jarrow et al. (2004). The discount function satisfies $D(t) > 0$ and $D(0) = 1$ while $y(t)$ and $f(t)$ are both positive functions. In this manuscript we will refer to the term structure using any of these equivalent representations.

In order to estimate the term structure, the discounted cash flow (DCF) can be used to link bond prices to the discount curve. A bond is a debt in favor of the bondholder, who receives in return a cash flow composed of interest(coupon) and the payment of the principal at the set maturity date. The DCF principle states that an investor is willing to pay for a given bond, b, the sum of the present value of the

remaining payments in the cash flow:

$$P_{DCF,b} = \sum_{j=1}^{m_b} \mathbf{CF}_b(j) * D_b(t_{b,j}), \tag{2.2}$$

where $P_{DCF,b}$ denotes the DCF bond price, $\mathbf{CF}_b$ is the cash flow vector including the m_b remaining payments, and $D_b(\cdot)$ is the discount curve of the bond evaluated at the time $t_{b,j}$ when the jth cash flow is paid. If needed, the discount function can be written in terms of $y(t)$ or $f(t)$ using equation (2.1). The discount function reflects the time value of money as well as a risk premium.

Based on the DCF principle, we can estimate the term structure as follows. First, any of the equivalent representations of the term structure are approximated using a parametric function with vector of parameters $\boldsymbol{\theta}$. Such a parametric function is called an approximating function. Next, the discount function is written in terms of the approximating function by using equation (2.1). Then, the discount function is used to compute the DCF bond price which now is a function of $\boldsymbol{\theta}$. And, finally, $\boldsymbol{\theta}$ is estimated by comparing the DCF price to the observed price of each bond in the sample; observed prices are equal to the quoted flat price plus accrued interest. The basic estimation problem is to find a discount curve with optimal explanatory power, that is, a discount curve that minimizes pricing errors with respect to a given norm. For example, using a quadratic loss function, the estimated term structure corresponds to $\boldsymbol{\theta}$ that minimizes

$$L(\boldsymbol{\theta}) = \sum_b \omega_b \left([P_b + ai_b] - P_{DCF,b}(\boldsymbol{\theta})\right)^2, \tag{2.3}$$

where P_b is the quoted flat price of the bond b, ai_b denotes the accrued interest, and each bond weight, ω_b, can be set based, for example, on the duration of the bond.

Several functional forms have been proposed as approximating functions to estimate the term structure. Two popular alternatives are splines (McCulloch 1971) and exponential polynomials (Nelson and Siegel 1987). Ioannides (2003) used in-sample and out-of-sample tests to compare the performance of these functional forms and found that parsimonious representations based on exponential polynomials perform better than those based on splines because the latter tend to overfit the data. In addition to this empirical evidence, we use exponential polynomials in the proposed framework (see Section 2.2) because they are parsimonious representations of the term structure. In particular, the functional form introduced by Nelson and Siegel (1987) for the yield curve is a four-parameter representation given by

$$y(t) = \beta_0 * \mathbf{1} + \beta_1 \left(\frac{1 - \exp(-t/\gamma)}{t/\gamma} \right) + \beta_2 \left(\frac{1 - \exp(-t/\gamma)}{t/\gamma} - \exp\left(-\frac{t}{\gamma} \right) \right). \quad (2.4)$$

Since $\beta_0 = y(\infty) > 0$ and $\beta_0 + \beta_1 = y(0) > 0$, it follows that $\beta_0 > 0$ and $\beta_1 > -\beta_0$. In addition, β_2 controls the shape of the term structure, which can include humps, S, and monotonic curves. And finally, $\gamma > 0$ determines how fast (slow) the loadings decay to zero. The three loadings in (2.4) have the following interpretation, respectively. The first one is a constant, hence it describes the behavior of the yield curve in the long term. The second is equal to one at zero and decays when the maturity increases, therefore it corresponds to the short term. Finally, the third loading corresponds to the medium term because as t increases, the loading, which is equal to zero at $t = 0$,

first increases and then decays to zero.

We use the rest of this section to discuss estimation methods for corporate bonds. Their term structure is commonly estimated by first grouping the bonds by credit rating levels, and then independently estimate the term structure of each class. (Schwartz 1998). We follow Houweling et al. (2001) and refer to this two-stage procedure as a single-curve approach; such a name emphasizes that the estimation is performed independently for each group. The use of credit ratings for determining groups of bonds is a limitation for this approach because empirical evidence suggests that other bond characteristics, in addition to bond ratings, influence the term structure of corporate bonds (Elton et al. 2004). A logical procedure to include such factors would be to use them for determining finer classifications of bonds. However, such an alternative is not feasible given the current estimation methods because, as Elton et al. (2004) pointed out, this would result in classifications with too few bonds within each group to estimate term structures with any accuracy.

A natural criterion to group corporate bonds is by issuer company. Such classification is relevant because the resulting estimators approximate the term structure of individual firms, and consequently, they reflect the uniqueness of a firm's credit risk. Obtaining such estimators is challenging. Most companies only issue a handful of bonds. Jarrow et al. (2004) developed the first model for the term structure of individuals firms. They proposed to model the term structure of corporate debt by adding a spread to the term structure of government bonds. A Bayesian version of

this model was introduced by Li and Yu (2005). Jarrow et al. (2004) and Li and Yu (2005) applied their method to a case study consisting of bonds issued by AT&T with information over the 21-month period of April 1994 to December 1995. On average, 4.3 bonds were used in each month to fit the term structure of AT&T. However, they did not include in-sample and out-of-sample tests to compare the performance of their estimators with those produced with the single-curve approach. Furthermore, it is not clear from the case study that this model has wide applicability to other firms. Specifically, their case study does not include bonds with low credit rating levels for which the corresponding term structure separates greatly from the risk-free term structure. Finally, the data set described in Section 2.3, which includes current information for 2009, shows an average number of bonds per company equal to 2.1, which is half the size of the average number of bonds in the case study presented by Jarrow et al. (2004) and Li and Yu (2005).

Estimation methods able to produce accurate estimators based on small samples of bonds are therefore needed because they can be used to estimate term structures for groups in fine classification, in particular, term structures of individual firms. We introduce a model an estimation strategy that directly addresses this issue. The key feature of our approach is to compensate the small number of bonds by jointly modeling multiple term structures so that we can pool information and borrow strength across them. Unlike Jarrow et al. (2004), we do not focus on the relationship between corporate and government term structures. Instead we propose to take advantage of

the information shared across all the corporate term structures. The details of our model appear in the following section.

2.2 A Semi-parametric Bayesian Population Model

To jointly estimate n term structures, we propose a Bayesian model whose hierarchical structure resembles a Bayesian population model. If the vector of parameters characterizing the ith term structure is denoted as $\boldsymbol{\theta}_i$ and P_{ib} is equal to the logarithm of the price of the bth bond corresponding to the ith term structure, then the Bayesian hierarchical model we are proposing includes three main components:

$$p(P_{ib}|\boldsymbol{\theta}_i), \qquad p(\boldsymbol{\theta}_i|\phi), \qquad p(\phi), \tag{2.5}$$

where $p(P_{ib}|\boldsymbol{\theta}_i)$ links bond prices and term structures via a non-linear regression model, $p(\boldsymbol{\theta}_i|\phi)$ is the prior for the vector of parameters $\boldsymbol{\theta}_i$, and $p(\phi)$ denotes the probability model of the hyperparameters. The hierarchical structure given by equation (2.5) corresponds to a population model from a Bayesian perspective. Population models are widely used in some disciplines. See, for example, Rosner and Müller (1997) and references therein for applications in pharmacokinetic studies. In spite of the popularity of this modeling approach, our proposed model is, to the best of our knowledge, the first one that adapts such an approach to estimate the term structure of interest rates.

The rest of this section describes the distributional assumptions in our model for each component in (2.5). We also explain how to introduce bond weights into the

model.

2.2.1 Non-linear Regression Model for the Discount Curve

We use the discounted cash flow (DCF) approach to link bond prices to the discount curve. Specifically, using the indexes ib to denote the data of the bond b with term structure i, the bond prices are modeled as

$$P_{ib} = \Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}) + \epsilon_{ib}, \tag{2.6}$$

where $\Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})$ is equal to DCF bond price computed with the cash flow vector $\mathbf{CF}_{ib}$ (see equation (2.2)), and ϵ_{ib} is an error term. The use of an error term is necessary because the exact equality between observed and DCF prices does not hold in practice due to market imperfections (Bliss 1997; Houweling et al. 2001). Usually, the error terms in (2.6) are assumed to be normally distributed, which in turn implies that the prices are normally distributed with mean $\Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})$. In our model, however, we assume that the prices follow a t Location-Scale distribution given by

$$P_{ib} \sim t_\nu \left(\Psi \left(\boldsymbol{\theta}_i, \mathbf{CF}_{ib} \right), \sigma^2 \right), \tag{2.7}$$

where ν denotes the degrees of freedom, the location parameter is equal to $\Psi \left(\boldsymbol{\theta}_i, \mathbf{CF}_{ib} \right)$ and the scale parameter is σ^2 (Gelman et al. 2004). We use a t-distribution because it allows for greater deviations between observed and theoretical prices than a normal distribution. Finally, the t distribution in (2.7) is equivalent to the following mixture

of normal distributions

$$\begin{aligned} P_{ib} &\sim N\left(\Psi\left(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}\right), V_i\right) \\ V_i &\sim \text{Inv} - \chi^2(\nu, \sigma^2), \end{aligned} \tag{2.8}$$

where $\text{Inv} - \chi^2(\nu, \sigma^2)$ denotes a scaled inverse-chi-squared distribution with mean $(\nu/(\nu-2))\sigma^2$ (Gelman et al. 2004). The equivalent representation in (2.8) is used because it facilitates the introduction of weights into the model and the design of a sampling algorithm for the proposed model.

To complete the specification of $\Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})$, we set the approximating function of the yield curve as

$$\begin{aligned} y(t,\ [\beta_0, \alpha, \beta_2, \gamma]) = \beta_0 \left(1 - \frac{1-\exp(-t/\gamma)}{t/\gamma}\right) + \alpha \left(\frac{1-\exp(-t/\gamma)}{t/\gamma}\right) \\ + \beta_2 \left(\frac{1-\exp(-t/\gamma)}{t/\gamma} - \exp\left(-\frac{t}{\gamma}\right)\right), \end{aligned} \tag{2.9}$$

where $t > 0$ denotes time, and the parameters satisfy $\beta_0 > 0$, $\alpha > 0$, $\gamma > 0$ and $\beta_2 \in \mathbf{R}$. The approximating function (2.9) and the functional form introduced by Nelson and Siegel (1987) are equivalent, but in the former the only condition on the parameters, if any, is to be strictly positive. Because of that feature, we can compute the logarithm of the parameters β, α and γ and use the parameterization $\boldsymbol{\theta} = [k_0 \log(\beta_0),\ k_0 \log(\alpha),\ 10\, k_0\, \beta_2,\ k_0 \log(\gamma)]$, where k_0 is a positive integer set to produce numerical stability. This parameterization is favored because the coordinates of $\boldsymbol{\theta}$ have no restrictions on the values they can take. Such a feature is necessary because, as described in Section 2.2.2, we use a mixture of multivariate normal distributions to model the parameters of the yield curve. The integer k_0 increases the scale of the coordinates of $\boldsymbol{\theta}$. Such an increase is needed for numerical stability. Specifi-

cally, when $k_0 = 1$ the covariance matrix of $\boldsymbol{\theta}$ shows a small determinant that leads to numerical errors when modeling its inverse (see the hyperprior for $\mathbf{S}^{-1}$ in Section (2.2.2)). In our experience, a value of $k_0 = 50$ is adequate to avoid the numerical problem described above.

The approximating function given in equation (2.9) is used to describe each one of the n term structures being estimated. Therefore, each term structure is characterized by a four-dimensional vector $\boldsymbol{\theta}_i$, for $i = 1, \ldots, n$.

2.2.2 Prior Distribution: Dirichlet Process Mixture

In order to produce reliable estimators based on small samples, we propose to jointly model the n individual regression models defined in Section 2.2.1 so that we can borrow strength across them. This intuitive idea translates into a Hierarchical model (2.5) in which the parameters $\boldsymbol{\theta}_i$ share the same population distribution. Such common population distribution need to be flexible enough so that it can reflect the heterogeneity in the underlying population. We propose to use a mixture prior whose features are described below.

We model the interindividual variation $p(\boldsymbol{\theta}_i|\phi)$ with a mixture of normals with weights w_h, locations μ_h, and common covariance matrix $\mathbf{S}$.

$$\boldsymbol{\theta}_i \overset{iid}{\sim} M(\boldsymbol{\theta}) \qquad \text{with} \qquad M(\boldsymbol{\theta}) = \sum_{h=1}^{\infty} w_h N\left(\boldsymbol{\mu}_h, \mathbf{S}\right). \tag{2.10}$$

Although the mixture in equation (2.10) is infinite, the hyperprior that we introduce below implies that most of the weight is assigned to only a few components. The use of

normal distributions in the mixture allows computationally efficient implementation of the full posterior inference. A common covariance matrix across the components is assumed because it allows us to reduce the number of parameters in the model. Finally, we set the prior of $\boldsymbol{\theta}_i$ as a mixture so that our model can accommodate the heterogeneity in the population such as outliers, over-dispersion, multiple modes and skewness. In particular, a mixture prior down weights the influence of outliers by assigning them to a component in the mixture, and thus, limiting the effect on the mixture components modeling those bonds with a typical behavior. Outliers can appear, for example, if companies are digressing to junk status before their ratings change.

The mixture in equation (2.10) is equivalent to

$$\begin{aligned} \boldsymbol{\theta}_i &\sim N(\boldsymbol{\mu}_i, \mathbf{S}) \\ \boldsymbol{\mu}_i &\sim G = \textstyle\sum_{h=1}^{\infty} w_h \delta\left(\boldsymbol{\mu}_h\right), \end{aligned} \tag{2.11}$$

where the function $\delta\left(x\right)$ assigns probability 1 to the value of x and 0 elsewhere. With the notation in equation (2.11), the parameters of the prior mixture are written as G, S, where G is a discrete distribution on $\boldsymbol{\mu}$ with possible values $\boldsymbol{\mu}_h$ and probabilities w_h, for $h = 1, \ldots, \infty$.

Because of the lack of information about the underlying distribution of $\boldsymbol{\theta}_i$, we treat $\{G, \mathbf{S}\}$ as random so that the weights and moments of each component are data driven, including the number of components with practically significant weights. Specifically, we model G as a random measure generated from a Dirichlet Process

(DP) with base measure G_0 and total mass parameter M, that is, $G \sim \text{DP}(G_0, M)$. The mean of the random measure G is given by G_0, while M is a scaling factor that determines the variance of G around G_0 (Ferguson 1973). The use of DP in mixture priors is referred in the literature as DP mixture (DPM) priors. For further discussion see, for example, Escobar and West (1995). Regarding $\mathbf{S}$, we adopt the usual conjugate inverse Wishart prior $\mathbf{S}^{-1} \sim Wishart(r, (r\mathbf{R})^{-1})$ with r degrees of freedom and mean $r(r\mathbf{R})^{-1} = \mathbf{R}^{-1}$. The mixing measure as well as the covariance matrix are common to all the parameters $\boldsymbol{\theta}_i$. Thus, the posterior inference will take advantage of the information shared across the term structures.

To complete our model we specify a hyperprior on $\{M, G_0\}$. We model the uncertainty on $\{M, G_0\}$ in order to reduce the chance of affecting the posterior results due to an inappropriate selection if they were considered non-random. However, this approach increases the complexity of the model. To reach a middle point between flexibility and complexity, we use hyperpriors that allow for an efficient implementation of the model. Specifically, M is given a gamma distribution and G_0 a multivariate normal: $M \sim Ga(a_m, b_m)$ and $G_0 \sim N(\mathbf{b}, \mathbf{B})$. The moments $\mathbf{b}$ and $\mathbf{B}$ are chosen to be conjugate to the kernel of the mixture random effects model: $\mathbf{b} \sim N(\mathbf{b}_0, \mathbf{B}_0)$ and $\mathbf{B}^{-1} \sim Wishart\,(w, (w\mathbf{W})^{-1})$.

Finally, we consider the distributional assumptions for the parameters ν and σ^2 at the top level of the hierarchical model (see equation (2.7)). For simplicity, the number of degrees of freedom ν is considered fixed while the scale parameter follows

a gamma distribution, $\sigma^2 \sim Ga(a_\tau, b_\tau)$. Different values for ν will be considered to evaluate its impact on the estimators produced by our model.

2.2.3 Weights

We incorporate bond weights into our model by changing the variance of the error terms in (2.8) as follows:

$$P_{ib} \sim N\left(\Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}), V_i(\omega_{ib})^{-1}\right), \tag{2.12}$$

where ω_{ib} is the weight of the bond b corresponding to the term structure i. With this approach, the effect of the weights is similar to that in equation (2.3). For each term structure i, maximizing the induced likelihood is equivalent to minimizing $\sum_b \omega_{ib} \left(P_{ib} - \Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})\right)^2$. We define the bond weights as

$$w_{ib} = \frac{\frac{1}{d_{ib}}}{\sum_b \frac{1}{d_{ib}}},$$

where d_{ib} is equal to the Macaulay duration of the ibth bond. The weights corresponding to the same term structure add up to one, that is, $\sum_b w_{ib} = 1$.

2.2.4 Posterior Inference

The posterior distribution of the model described in this section does not have a closed form. Therefore, to sample from the posterior distribution we use a Markov chain Monte Carlo (MCMC) scheme. Such sampling scheme can be efficiently implemented since both the kernel of the mixture and the base measure G_0 are normally distributed

(MacEachern and Müller 1998). In our sampling scheme we resample $\boldsymbol{\theta}_i$ by using the adaptive Metropolis (AM) algorithm introduced by Haario et al. (2001). A complete description of the sampling scheme is available in the Appendix A.

2.3 Empirical Application

Our model does not use specific characteristics of government and/or corporate bonds. Therefore, it can be used to jointly estimate any combination of corporate and/or government term structures. In this section, however, we illustrate the performance of our new approach by using a data set composed entirely of U.S. corporate bonds. We focus on corporate bonds to demonstrate the ability of our new framework to produce accurate estimators based on small samples of bonds. This section describes the data set we use, gives details about the implementation of the estimation methods that are being compared, and reports results for in-sample and out-of-sample tests.

The bond data used in this section were obtained by combining information from two data bases: the Trade Reporting and Compliance Engine (TRACE) introduced by the Financial Industry Regulatory Authority, and The Mergent Fixed Income Securities Database (FISD) for academia. Both databases were accessed through the Wharton Research Data Services (`http://wrds.wharton.upenn.edu/`). The database TRACE, introduced in July of 2002, consolidates transaction data on 100 percent of over-the-counter activity representing over 99 percent of total U.S. corporate bond market activity in over 30,000 securities. Using TRACE we can obtain,

for a given trading day, a list of the bonds traded and their prices. However, other information about those bonds (time to maturity, coupon, payment frequency, issuer, etc.) are not available in TRACE. We obtained such information in Mergent-FISD, a comprehensive database of publicly-offered U.S. bonds that provides details on debt issues and the issuers on over 140,000 securities.

For illustration, we consider the U.S. corporate bonds traded on June 15, 2009. For each transaction, our data set includes price, time, yield or effective rate of return, and volume. The characteristics per bond in our data set include issuer company, maturity date, coupon, face value, payment frequency, clean prices, and Moody's credit ratings. We discard transactions that have an associated negative yield because that is a sign of poor liquidity and/or a data entry error. Only 129 out of the 42,626 transactions in the complete data set have a negative yield. We construct a sample of fixed coupon, non-callable, non-putable, investment grade bonds (AAA, AA, A, BBB), with maturity between 1 and 20 years. The bond prices are computed as the average across all the transactions included in the data set. Our final sample contains 599 bonds.

Two methods are compared in this section: our Bayesian model and the popular single-curve approach. Details about the implementation of each method are as follows.

Our Bayesian model includes an MCMC scheme to sample from the corresponding posterior distribution. Such implementation is written in the programming language

C. The parameters of the ith term structure are estimated as the posterior mean of the vector of parameters $\boldsymbol{\theta}_i$. The posterior mean is approximated by averaging the posterior sample. As explained in the description of the model, the number of degrees of freedom ν in the t distribution of the likelihood is not random. Hence, we use different values for ν to evaluate its effect on the performance of the model. Specifically, two values for ν are considered, 3 and 10. The case $\nu = 3$ corresponds to the heavier tails.

Regarding the single-curve method, it produces estimators by grouping the bonds based on credit rating level, and independently estimating the term structure of each class using the DCF principle. The functional form proposed by Nelson and Siegel (1987) is used as an approximating function of the discount function. The computations are performed using the package "termstrc," which is written in the R system for statistical computing (Ferstl and Hayden 2008). It has been proposed by Elton et al. (2004) to filter out from the sample those bonds with price residuals greater than 5 dollars.

We compare the performance of the estimation methods by measuring their in-sample goodness of fit. Specifically, we compare the price residuals of each model, where the price residual of a bond is equal to the market price minus the theoretical DCF bond price (see equation (2.2)) which is calculated using the estimated discount curve. Comparing price residuals is appropriate because term structure models should be able to explain market prices accurately since interest rates are the main

determinants of bond prices. To summarize the residuals we compute the Root Mean Squared Error (RMSE) given by

$$\text{RMSE} = \left(\frac{1}{n} \sum_{b=1}^{n} e_b^2 \right)^{1/2}, \tag{2.13}$$

where e_b denotes the price residual of bond b. The term structure model with the lower RMSE provides the better fit.

We use our Bayesian model to estimate the term structure of corporate bonds grouped by credit rating and by issuer company, respectively. In both cases, we compare our estimators with those produced by the single-curve method which groups the bonds by credit rating. Before showing the result of those comparisons, we remark how different the groups of bonds are in term of their sizes. When splitting the bonds by rating levels, there are four groups corresponding to the credit rating levels AAA, AA, A, and BBB. Each of those groups include 31, 117, 306 and 145 bonds, respectively. In contrast, grouping the bonds by issuer result in 197 groups, each one corresponding to a company, 114 (58%) of them including only one bond (see Table 2.1).

We first compare the estimated term structures of corporate bonds grouped by credit rating. The estimated yield curves by method appear in Figure 2.1. The yield curves produced with our Bayesian method show the expected relationship between credit risk and yield. That is, the lower the credit rating, the higher the yield. Identical results are obtained with 3 and 10 degrees of freedom. In contrast, the curves estimated using the single-curve method fail to show such a pattern for maturities

	Number of Bonds					
	1	2	3	4	5	≥ 6
# Companies	114	33	17	5	10	18
	(58%)	(17%)	(9%)	(3%)	(5%)	(9%)

Table 2.1 : Distribution of companies by number of bonds. The table refers to U.S. corporate bonds with information for June 15, 2009. Percentages do not add up to 100% due to rounding

higher than 10 years, despite the fact that our data set includes bonds with maturities up to 20 years. We conclude that our data-driven model is adequate to identify the relationships between the term structures being estimated. Finally, besides the differences among the estimated curves by method, in terms of in-sample goodness of fit both models show similar results; the RMSEs of the single-curve method and our Bayesian model (with $\nu = 3$) are 8.4 and 8.5, respectively.

We show later in this section that more accurate estimators can be produced when estimating the term structure by company rather than by rating class. This relative performance suggests that the resulting groups are heterogeneous, and hence, rating levels are not a sufficient metric to split corporate bonds for the purpose of estimating their term structure. This raises the issue of understanding what factors are responsible for the heterogeneity among bonds with the same rating level. Although such an aspect is not a primary goals of our research, we make the following comments on that direction. An empirical study described in Elton et al. (2004) found that

the following five factors are important: the finer rating level categories introduce by rating agencies when combined with maturity measures; differences between rating levels given by different rating agencies; differences between company and bond rating levels; the coupon size of the bonds; and finally, the age of the bond. In addition, other factors that may decrease the ability of rating levels to define homogeneous groups are the age of the ratings as well as the volume of the transactions.

We now use the proposed Bayesian population model to estimate the term structure of individual corporations. Specifically, we jointly estimate the term structure of 197 individual firms. The estimated yield curves with $\nu = 3$ and $\nu = 10$, respectively, are shown in Figure 2.2. In general, there are some visual differences among the estimated curves corresponding to each value of ν. However, we will show that those estimators have very similar performance in terms of in-sample and out-of-sample tests.

In-sample tests that verify the quality of the estimators produced with our model are discussed below. We use as a basis for comparison the estimators produced by the single-curve approach. The absolute price residuals of our Bayesian population model are smaller than those produced with the single-curve method (see Figure 2.3). Table 2.2 includes summary statistics of the absolute price residuals by model. Comparing the means, our model provides a noteworthy reduction of 78%. Similar results are found by comparing the RMSE of each method. Specifically, the single-curve method has a RMSE equal to 8.37 while the RMSE of our Bayesian model is

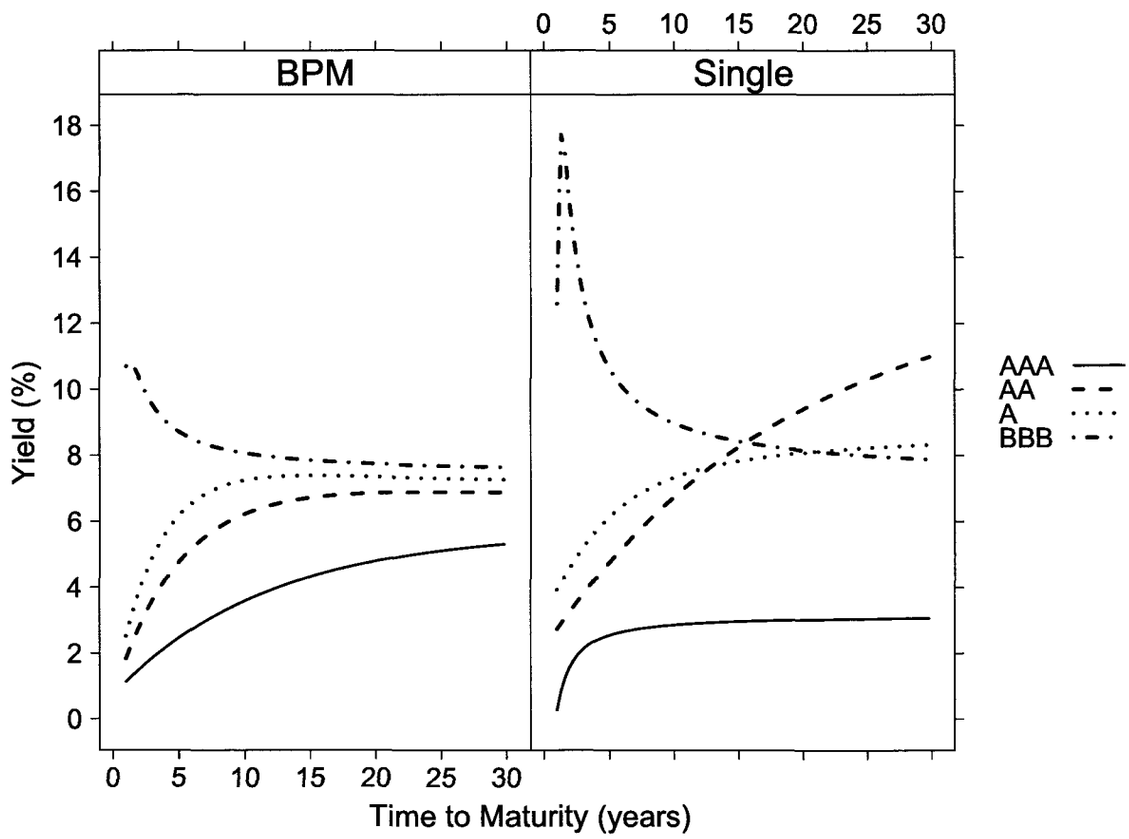
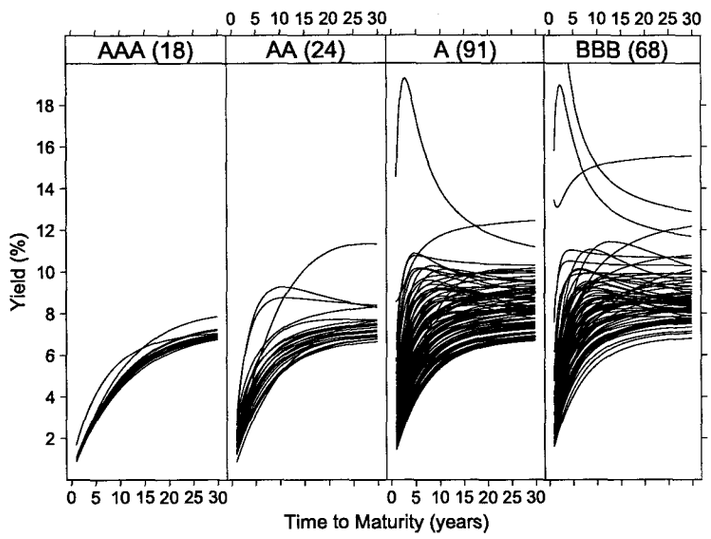
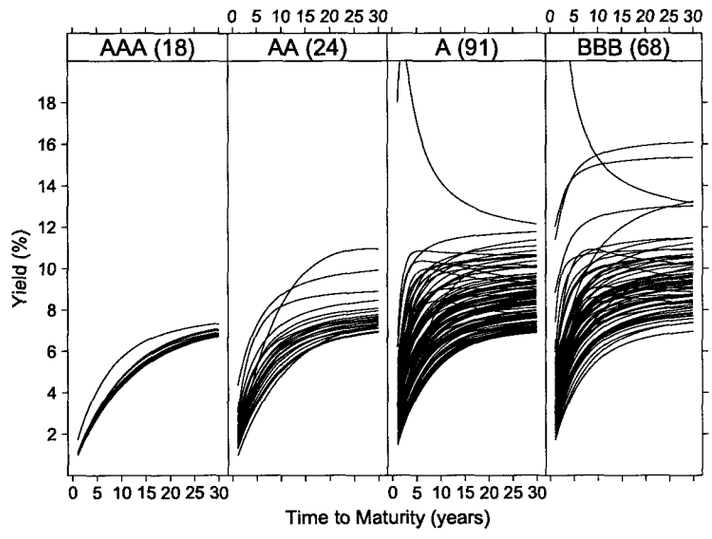


Figure 2.1 : Yield curves for AAA, AA, A and BBB bonds, respectively. The estimators produced with our Bayesian population model (BPM), with $\nu = 3$, are in line with the theory in terms of their "order." The estimated yield curves obtained with $\nu = 10$ are not shown because they are identical to those produced with $\nu = 3$. In contrast, the single-curve estimators (Single) cross each other.



(a) $\nu = 3$



(b) $\nu = 10$

Figure 2.2 : Yield curves of individual firms. The estimators were produced using our Bayesian Population model with ν equal to 3 and 10. Each panel corresponds to a given credit rating and includes the yield curves for companies with at least one bond having such a rating level. The number in parenthesis next to the rating level is equal to the number of curves being displayed. Since only three companies in our data set include bonds with different rating levels, only three yield curves appear more than once.

only 2.24, a reduction of 73%. Figure 2.3 also shows that the absolute price residual for the single-curve approach are higher for bonds with low credit rating level, hence showing that it is not reasonable to assume that bonds with the same credit rating level share similar term structures. An interpretation of the results described above follows from considering the well-known decomposition of the mean squared error into bias and variance. The good performance of the proposed model is the result of dramatically reducing the bias by approximating the term structure of corporate bonds with the term structure of the issuer company rather than using estimators by rating class.

Method	Minimum	Q1	Median	Mean	Q3	Maximum
BPM(3)	0.00	0.23	0.62	1.14	1.58	18.16
BPM(10)	0.00	0.25	0.59	1.14	1.52	18.18
Single	0.01	0.93	2.74	5.24	6.75	42.24
Percentage Difference	-50%	-73%	-78%	-78%	-77%	-57%

Table 2.2 : Summary statistics for absolute price residuals by model. Q1 and Q3 denote the first and third quartiles, respectively. Compared with the single-curve estimators (Single) obtained for each credit risk class, the estimators of the term structure of individual firms computed with our Bayesian population model, BPM(ν), provide a noteworthy reduction of 78% in the mean of absolute price residuals. The percentage differences compare the statistics of BPM(10) and Single.

To compare term structure estimation methods, it is good practice to use out-of-sample measures in addition to in-sample goodness of fit tests (Bliss 1997). We

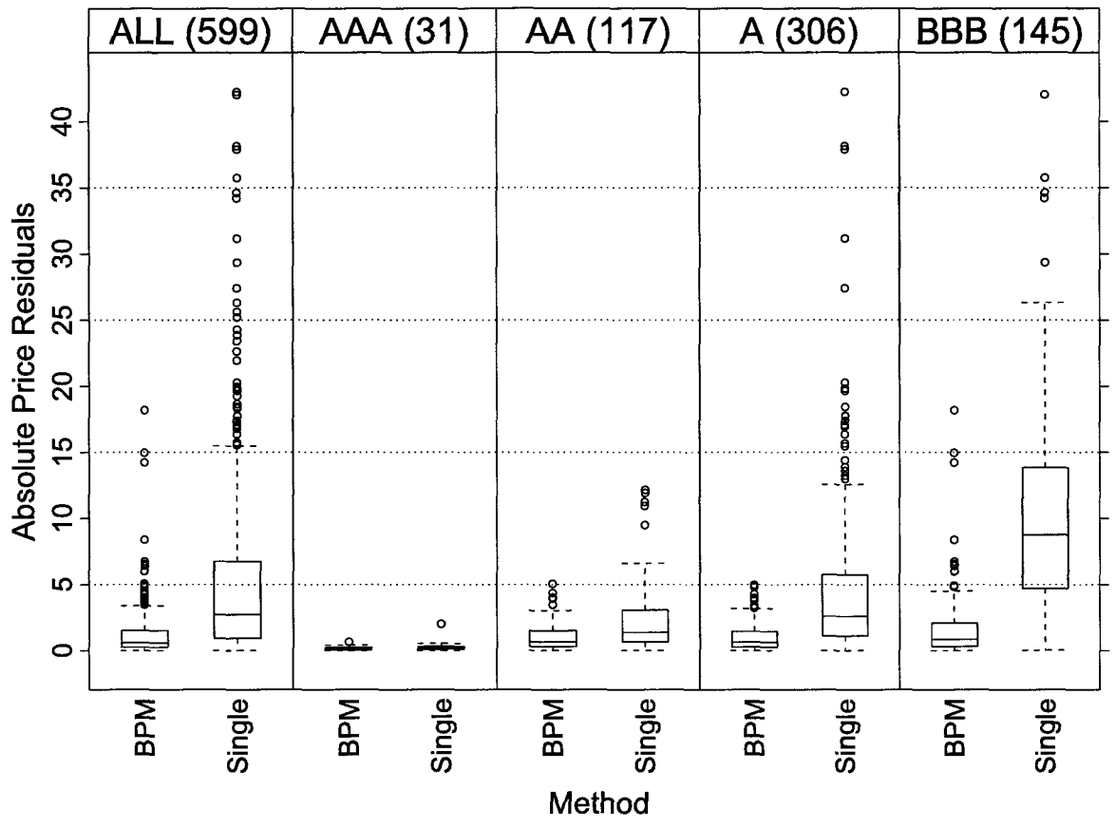


Figure 2.3 : Boxplots for in-sample absolute price residuals arranged by method and credit rating. The labels "AAA", "AA","A" ,"BBB" refer to credit rating while "ALL" indicates that all the residuals are being considered. The number in parenthesis is equal to the number of bonds in each class. Using our Bayesian population model (BPM) with $\nu = 10$ to estimate the term structure of individual firms results in smaller absolute price errors than those produced by the traditional single-curve model (Single) which groups the bonds based on credit rating. Boxplots for price residuals obtained when using the BPM model with $\nu = 3$ are not shown, because they are almost identical to those produced with $\nu = 10$

use cross-validation to obtain out-of-sample measures. One round of cross-validation starts by partitioning the dataset into complementary subsets: a training set and a test set. The training set is used to fit the term structures, then for each bond in the test set we use the corresponding fitted model to compute its theoretical DCF price. Finally, we compute residuals by subtracting the DCF price from the market prices. To summarize the residuals, or prediction errors, we consider the following statistics:

$$\text{RMSPE} = \frac{1}{n_{test}} \sum_{b=1}^{n_{test}} e_b^2 \qquad \text{and} \qquad \text{MAPE} = \frac{1}{n_{test}} \sum_{b=1}^{n_{test}} |e_b|, \tag{2.14}$$

where RMSPE denotes the Root Mean Square Prediction Error, MAPE stands for the Mean Absolute Prediction error, the sum is performed over the bonds in the test set, the size of the test set is denoted as n_{test}, and e_b is the residual for the bond b. The statistics defined above correspond to one round of cross-validation. Multiple rounds using different partitions can be used to reduce variability; the statistics are averaged over the rounds.

Out-of-sample measures are used to evaluate the performance of our method when estimating the term structure of individual firms. The criteria we use to define training and test sets are explained below. First, we consider partitions where the test set is defined by selecting one bond from any company having exactly two bonds, while the training set includes the rest of the bonds in the data set. In other words, any company with exactly two bonds in the complete data set will become a one-bond company in the training set. Such partitions are of interest because, by using the bonds in the test set, we will be able to verify the reliability of our term structure

model when producing estimators of groups containing only one bond. We consider two such partitions. For any group consisting of exactly two bonds we order the bonds in increasing order by maturity. The first partition is obtained by including in the test set the first bond, while for the second partition the second bond is selected. Following similar ideas, we define other partitions. We identify all the companies with exactly m bonds and for each one of them we order its bonds with respect to maturity. We then move the kth bond of each company, $k \leq m$, into the test set. The rest of the bonds in the data set are kept in the training set. We repeat this procedure with $m = 3$ and 4. In total, we define 9 partitions corresponding to the combinations of positive integers (m, k), with m in $\{2, 3, 4\}$ and $k \leq m$.

We compute the out-of-sample measures for the estimators obtained with our model and, as we did with in-sample tests, we use the single-curve approach as a basis for comparison. The bonds in the training sets are grouped by different criteria depending on the estimation method; credit ratings for the single-curve method, and issuer company for our Bayesian model. We compute the price residuals of the bonds in the test set. For both values of ν, 3 and 10, the proposed Bayesian population model produces smaller absolute price errors than the single-curve method (see Figures 2.4 and 2.5). We summarize the residuals by computing the statistics defined in equation (2.14) and averaging the results of partitions with the same m in (m, k). Table 2.3 includes the statistics by method and show the better performance of our Bayesian model with $\nu = 10$. Specifically, we obtain a reduction of at least 52% in the MAPE

for any value of m. The results in 2.3 also show that with the proposed method, the out-of-sample results are better with $\nu = 10$ than when using $\nu = 3$. In particular, only with $\nu = 10$ all the price errors are smaller than 15 (see Figures 2.4 and 2.5).

Therefore, lighter tails in the distribution of the prices seems to provide a better fit. See Appendix C for comments on the performance of our term structure model under normally distributed prices.

	RMSPE			MAPE		
	(2)	(3)	(4)	(2)	(3)	(4)
BPM(3)	4.17	6.41	2.69	2.71	3.83	1.64
BPM(10)	3.30	3.81	1.92	2.50	2.65	1.33
Single	7.39	8.53	9.07	5.24	6.60	6.55
Percentage Difference	-55%	-55%	-79%	-52%	-60%	-80%

Table 2.3 : Root mean squared prediction error (RMSPE) and mean absolute prediction error (MAPE) by method. The table shows the average of the statistics RMSPE and MAPE over partitions with (m, k) having the same m (number in parenthesis). See text for a description of the partitions. In all cases the estimators of the term structures of individual firms produced with our Bayesian population model (BPM) outperforms those obtained with the single-curve approach (Single).

2.4 Simulation Example

In this section, we use a simulated data set to illustrate the performance of the proposed Bayesian population method when estimating the term structure of indi-

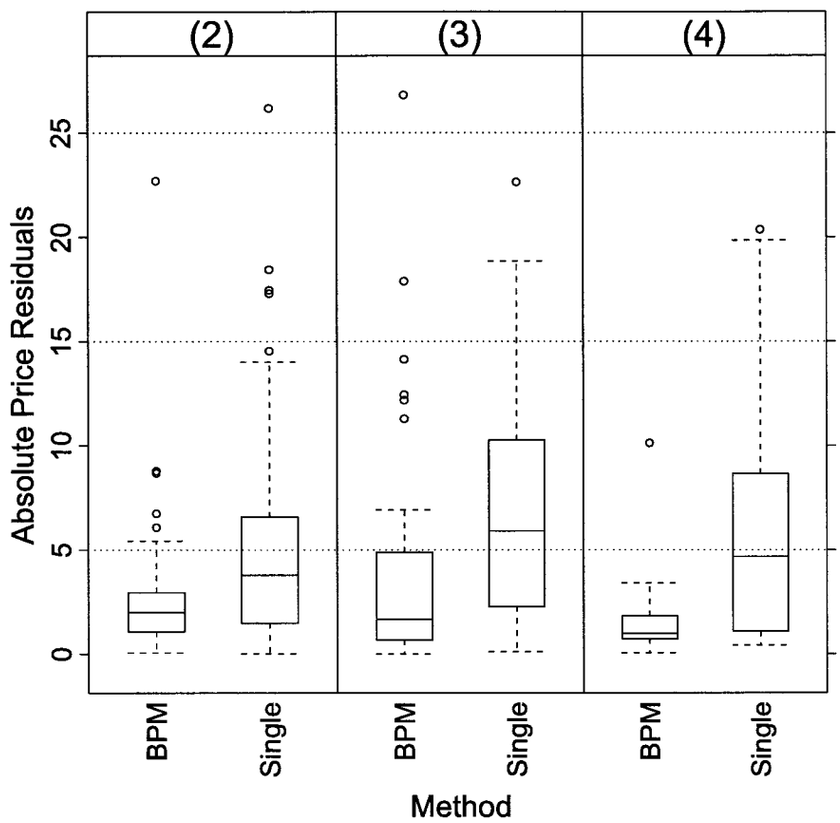


Figure 2.4 : Boxplots for out-of-sample absolute price residuals of bonds in test sets. The absolute residuals corresponding to the proposed Bayesian population model (BPM) with $\nu = 3$ are smaller than those produced with the single-curve method (Single). In the panel "(m)", for m = 2, 3, 4, each boxplot includes absolute price residuals of bonds belonging to the test set of any partition with combination (m, k) (see text for details on the partitions). Each boxplot in panels (2), (3) and (4) represent 66, 51 and 20 absolute residuals, respectively.

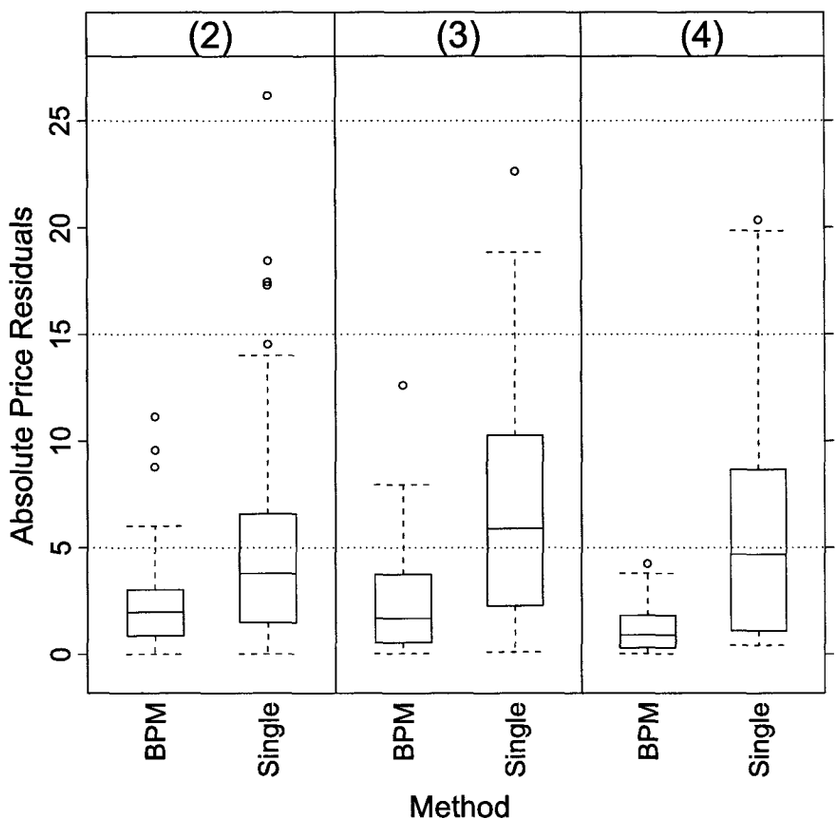


Figure 2.5 : Boxplots for out-of-sample absolute price residuals of bonds in test sets. The absolute residuals corresponding to the proposed Bayesian population model (BPM) with $\nu = 10$ are smaller than those produced with the single-curve method (Single). In the panel "(m)", for m = 2, 3, 4, each boxplot includes absolute price residuals of bonds belonging to the test set of any partition with combination (m, k) (see text for details on the partitions). Each boxplot in panels (2), (3) and (4) represent 66, 51 and 20 absolute residuals, respectively.

vidual firms. A simulated data set offers the advantage that the true underlying term structure of each company is known. Hence, instead of indirectly measuring the performance of an estimation method via bond prices, we can directly evaluate how accurately the estimators approximate the underlying term structure.

We use the following approach to generate a simulated data set. First, we use as a starting point the empirical sample described in Section 2.3 except bond prices. This implies that the resulting simulated data set will be identical to such an empirical data set with the exception of bond prices. Next, for each company, we generate its underlying term structure. Finally, the generated term structure is used to set the price of each company's outstanding bonds. The simulated data set includes 599 bonds corresponding to 197 companies and the cash flow of a given bond is the same in both the simulated data set and the empirical sample. An advantage of the approach just described is that it leads to realistic simulated data sets that reflect the heterogeneity found in empirical samples, which not only refers to the shape of each company's underlying term structure, but also to the characteristics of the bonds in the sample, for example, coupon size and time to maturity.

To implement the approach described above, we need to specify an algorithm to generate the underlying term structure of each one of the 197 companies as well as a procedure to simulate bond prices. In this section, we generate those term structures by drawing a sample from the posterior distribution that results of fitting the proposed Bayesian population model (with $\nu = 10$) to the empirical data described in Section

2.3. Specifically, the vectors of parameters, $\tilde{\boldsymbol{\theta}}_i$, for the underlying term structure of the companies in the sample were obtained by drawing a sample of size one from the joint posterior distribution of $\boldsymbol{\theta}_i$. Hence, we are parameterizing the underlying term structures with the functional form introduced by Nelson and Siegel (1987). A drawback of this algorithm is that it could induce a bias in favor of our modeling approach because we are using the output of the proposed model to generate the data. Ideally, we would like to generate the simulated term structures by using estimators produced by competing estimation methods. This option, however, is not feasible because, to the best of our knowledge, our framework is the first one able to estimate the term structure for companies with a small number of outstanding bonds.

Finally, we generated the bond prices in the simulated data set. Such prices need to be in agreement with the generated term structure given by $\tilde{\boldsymbol{\theta}}_i$. Using the discounted cash flow approach, the simulated bond price, $\tilde{P}_{ib}$, of the bth bond corresponding to the ith company is sampled from the model

$$P_{ib} \sim t_{\nu=10}\left(\Psi\left(\tilde{\boldsymbol{\theta}}_i, \mathbf{CF}_{ib}\right), \sigma^2 = 0.45\right).$$

The procedure just described allows to generate one simulated data set. We can, however, generate any number of simulated data sets which will differ in terms on the term structure of each company and the respective prices.

We generated 100 simulated data sets using the algorithm described above. (see Figure 2.6). For each simulated data set, we estimated the term structure of each company with both methods, the proposed Bayesian population model (with $\nu = 10$)

and the single-curve approach. When using the single-curve approach, the term structure of a given company is approximated with the term structure corresponding to the rating level of the outstanding bonds. Because we are using a simulated data set, we can directly compare estimated to true term structures. In particular, for each company, we computed yield residuals (the difference between the estimated and true yields) for maturities 2, 5, 10 and 20, respectively. The absolute yield residuals obtained with the proposed Bayesian population model are smaller than those obtained with the single-curve approach (see Figure 2.7). Across the 100 simulated data sets, the average percentage of companies with absolute yield residuals larger than 2% is less than 7.4% for the Bayesian model while for the single-curve can be as high as 45.29% (see Table 2.4).

2.5 Prediction Based on Covariates

The Bayesian population model described in Section 2.2 produces estimators based on the observed bond prices. However, there are other bond characteristics or covariates that are also related to the term structure. An advantage of introducing covariates into a term structure model is that we will be able to estimate the term structure that correspond to a given set of covariates. Such estimators could be used, for example, to price new issues of bonds for which no price data is available.

The structure of this section is as follows. We first extend the Bayesian population model described in Section 2.2 to include dependence on covariates. Then, we perform

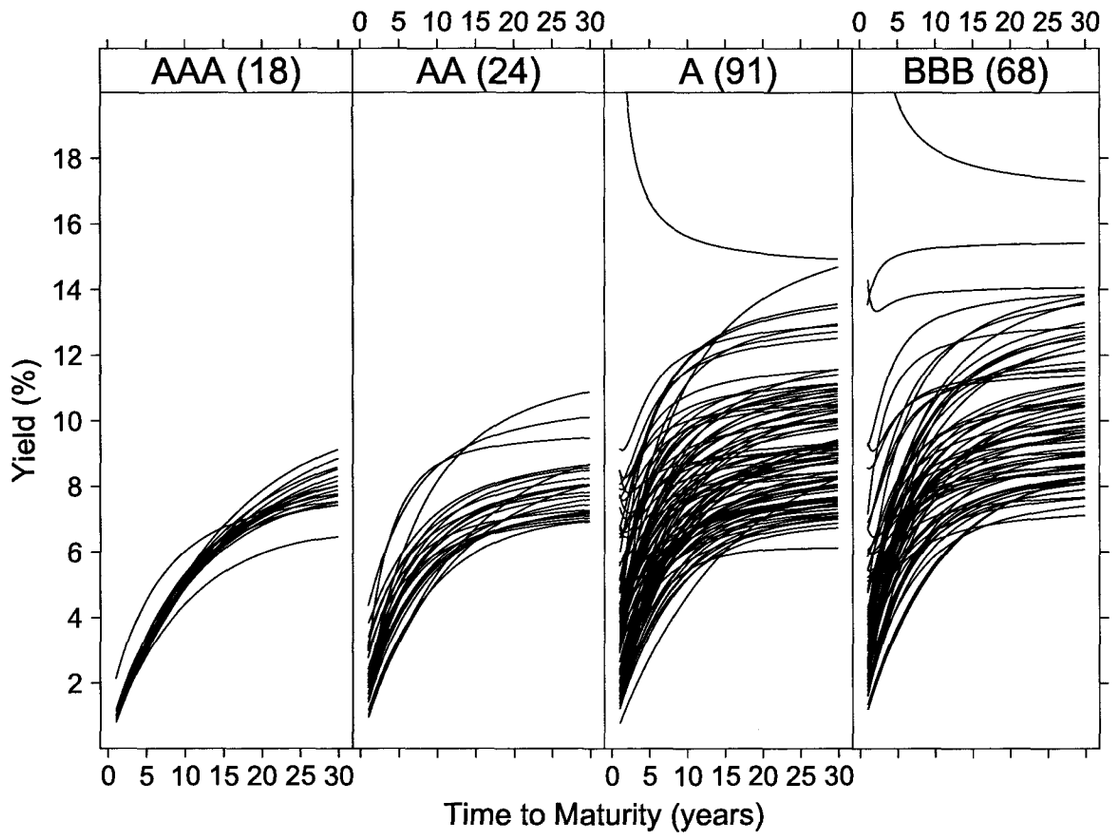


Figure 2.6 : Simulated yield curves of individual firms in one simulated data set. Each panel corresponds to a given credit rating and includes the yield curves for companies with at least one bond having such a rating level. The number in parenthesis next to the rating level is equal to the number of curves being displayed. Since only three companies in our data set include bonds with different rating levels, only three yield curves appear more than once. The curves displayed correspond to one simulated data set, however similar patterns are found in the 100 generated data sets.

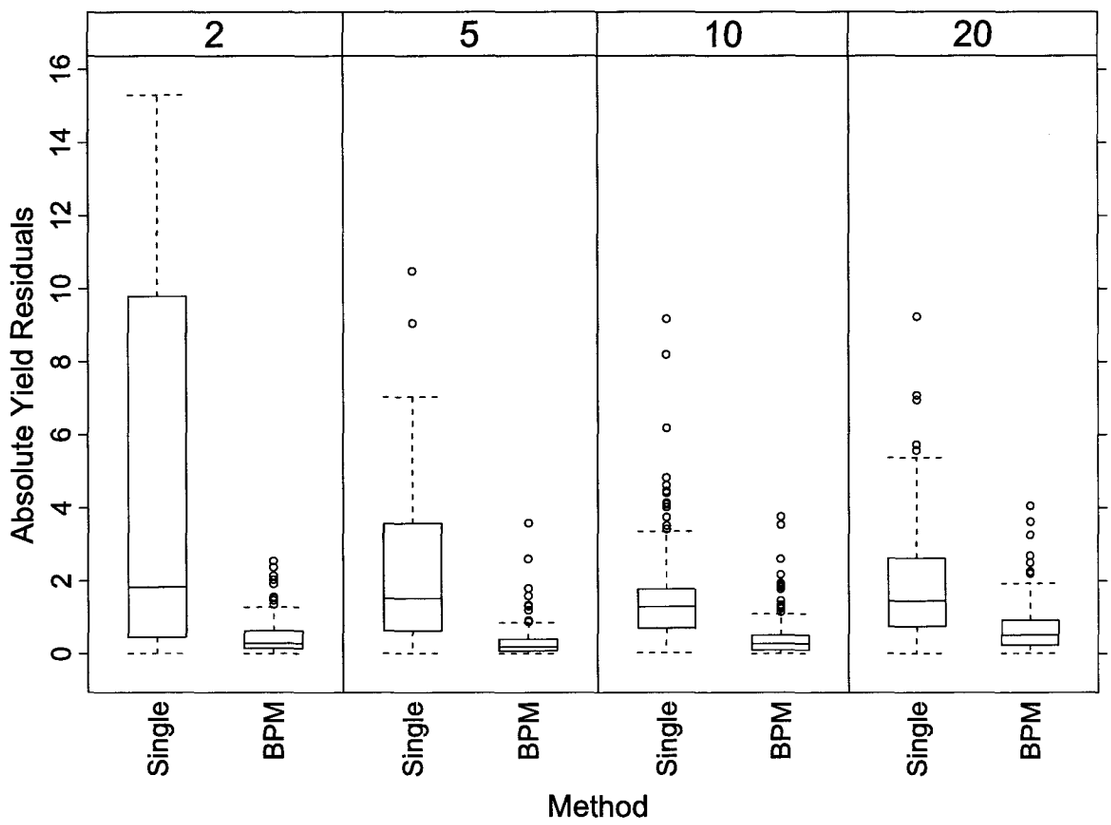


Figure 2.7 : Boxplots for absolute yield residuals by maturity and method for one simulated data set. Each panel refers to yield residuals for a different maturity; 2, 5, 10 and 20, respectively. The absolute yield residuals corresponding to the proposed Bayesian population model (BPM) with $\nu = 10$ are smaller than those produced with the single-curve method (Single). For summary statistics across the 100 simulated data sets, see Table 2.4.

	Mean				> 2 (%)			
	[2]	[5]	[10]	[20]	[2]	[5]	[10]	[20]
BPM	0.93	0.71	0.62	0.81	8.30	3.84	4.04	7.39
	(0.06)	(0.05)	(0.06)	(0.09)	(1.75)	(1.38)	(1.78)	(2.65)
Single	4.15	2.08	1.51	2.05	45.29	40.19	25.31	39.70
	(4.01)	(3.50)	(3.73)	(4.87)	(3.90)	(3.61)	(6.33)	(5.88)

Table 2.4 : Summary statistics for absolute yield residuals by maturity and method, average over 100 simulated data sets. The number in square brackets at the top of the table denotes the maturity of the yields being analyzed. The table shows, for each maturity, the mean of the absolute yield residuals (left) and the percentage of firms with absolute residual greater than 2 (right). For each method, the first row correspond to the average over the 100 simulated data sets while the second row shows in parenthesis the standard deviation. For all maturities, the yield residuals obtained with the proposed Bayesian population model (BPM) are smaller than those obtained with the single-curve approach (Single).

posterior inference using the sample of corporate bonds described in Section 2.3. Finally, we explain how the results in this section motivate the research problem we study in Chapter 3.

2.5.1 Extension of the Term Structure Model

Some of the bond covariates that influence their term structure include the economic sector of the issuer company, rating level, age, etc. Different approaches can be used in order to introduce such covariates into the proposed hierarchical model (2.5). We can, for example, modify the top level of the hierarchy by considering $p(P_i|\theta_i, x_i)$,

where x_i is the vector of covariates of the ith bond, or we can instead modify the second stage in the hierarchy with $p(\theta_i|x_i, \phi)$. Since $E[P_{ib}|\boldsymbol{\theta}_i]$ is equal to a nonlinear function, modifying the top level implies that the covariates are included into the nonlinear function or have them as offsets, that is, as additive terms. On the other hand, including the covariates into the second stage gives a flexible regression via DP. The modeling strategy described in this section follows the second flexible alternative. Specifically, we follow De Iorio et al. (2004) and model categorical covariates in an ANOVA fashion as described below.

For simplicity, suppose there is exactly one categorical covariate with three levels. The arguments below can be easily generalized for any number of covariates and levels. Let p be the dimension of $\boldsymbol{\theta}_i$, which in our current application is equal to 4. We replace the model in equation (2.11) by

$$\begin{aligned} \boldsymbol{\theta}_i &\sim N(\boldsymbol{\alpha}_i \boldsymbol{d}_i, \mathbf{S}) \\ \boldsymbol{\alpha}_i &\sim G = \textstyle\sum_{h=1}^{\infty} w_h \delta\left(\boldsymbol{\alpha}_h\right), \end{aligned} \tag{2.15}$$

where $\boldsymbol{\alpha}_i$ is a matrix with p rows and as many columns as the number of levels of the categorical covariate. With three levels, the matrix $\boldsymbol{\alpha}$ takes the form $\boldsymbol{\alpha} = [\boldsymbol{U}, \boldsymbol{V}_2, \boldsymbol{V}_3]$, where $\boldsymbol{U}$, $\boldsymbol{V}_2$, and $\boldsymbol{V}_3$ are column vectors. In addition, $\boldsymbol{d}_i$ is the design vector of the ith group which is equal to $(1,0,0)$, $(1,1,0)$, or $(1,0,1)$, when the observed value of the categorical covariate corresponds to the first, second, or third level, respectively. These modeling assumptions imply that when the categorical covariate is equal to the first level, the mean in the mixture is equal to the "overall mean" $\boldsymbol{U}$, while

for other cases, the mean in the mixture is equal to the "overall mean" $\boldsymbol{U}$ plus the corresponding "offset" of the categorical covariate $\boldsymbol{V}_k$, $k = 2, 3$.

We keep the same distributional assumption for G. That is, $G \sim DP(M, G_0)$, but now the base measure G_0 takes the form

$$G_0(\boldsymbol{\alpha}) = G_0^{\boldsymbol{U}}(\boldsymbol{U}) \times \Pi_{k=2}^{3} G_0^k(\boldsymbol{V}_k),$$

where $G_0^{\boldsymbol{U}} \sim N(\boldsymbol{b}, \boldsymbol{B})$ and $G_0^k \sim N(0, \boldsymbol{B}_1)$, $k = 2, 3$. That is, the prior distribution for the mean vector is multivariate normal with mean $\boldsymbol{b}$ and covariance matrix $\boldsymbol{B}$, while each offset vector has a multivariate normal prior with mean zero an covariance matrix $\boldsymbol{B}_1$.

In general, when there are d categorical covariates each having $c(\ell)$ levels, $\ell = 1 \cdots d$, the columns of $\boldsymbol{\alpha}$ include the overall mean vector $\boldsymbol{U}$, and $c(\ell) - 1$ offset vectors for each categorical covariate. Denoting the offset vectors as $\boldsymbol{V}_k^{\ell}$, $k = 2 \cdots c(\ell)$ and $\ell = 1 \cdots d$, the distribution of G_0 has the form

$$G_0(\boldsymbol{\alpha}) = G_0^{\boldsymbol{U}}(\boldsymbol{U}) \times \Pi_{k=2}^{c(1)} G_{0,1}^k(\boldsymbol{V}_k^1) \times \cdots \times \Pi_{k=2}^{c(d)} G_{0,d}^k(\boldsymbol{V}_k^d),$$

where $G_0^{\boldsymbol{U}} \sim N(\boldsymbol{b}, \boldsymbol{B})$ and $G_{0,\ell}^k \sim N(0, \boldsymbol{B}_\ell)$ for $k = 2 \cdots c(\ell)$ and $\ell = 1 \cdots d$. To complete the model, we set the hyperprior of the covariance matrices $\boldsymbol{B}_\ell$, $\ell = 1, \ldots, d$, to be

$$\boldsymbol{B}_\ell^{-1} \sim Wishart(w, (w, \boldsymbol{W})^{-1}).$$

The extension described above offers several advantages. It provides a flexible model where the effect of the covariates on the parameters does not need to be the

same for all the bonds. In fact, it depends on the columns of the matrix $\boldsymbol{\alpha}$ which changes for each component in the mixture. Furthermore, the implementation of this model is simple because its form is similar to the standard Dirichlet process mixture (DPM) model. See Appendix B for a general description of a MCMC sampling scheme. Further comments on introducing dependence on covariates into DPM models are discussed in Chapter 3.

2.5.2 Implementation

We illustrate the covariate-dependent model described in the previous section by using the sample of 599 bonds introduced in Section 2.3. As in Elton et al. (2004), we consider the following four covariates which affect the valuation of corporate bonds. First, the Moody's credit rating level which includes 10 levels: Aaa, Aa1, Aa2, Aa3, A1, A2, A3, Baa1, Baa2, and Baa3. Second, the industry group of the issuer which can take the values industrial and finance. Third, the differences between Moody's and Standard & Poor's credit ratings including three cases: S&P is higher than Moody's, Moody's is higher than S&P, and both ratings are equal. Finally, the fourth covariate is based on the age of the bond and it is equal to 1 if the bond is older than one year and 0 otherwise. The covariates above were available only for a subset of the 599 bonds in the sample. Hence, the analysis that follows is based on such a subset which includes 568 bonds.

For a given combination of the four covariates introduced above, the goal is to

estimate the term structure that is likely to be observed for bonds having such a combination of covariates. From a Bayesian perspective, such estimators can be produced by considering the predictive distribution of the vector of parameters $\boldsymbol{\theta}$ given the covariates of interests $\boldsymbol{x}_{new}$, that is, the predictive distribution $p(\boldsymbol{\theta}|\boldsymbol{x}_{new}, \text{Data})$. Details on how to get a sample from the predictive distribution can be found in Appendix B. For illustration, we produce such estimators considering four combinations of covariates. Such combinations differ in terms of the Moody's rating level, specifically, the levels Aaa, Aa2, A2 and Baa2 are considered. The other three covariates are the same for the four combinations. The industry group is finance, the Moody's and S&P credit ratings are the same, and the bond age is greater than one year. The pointwise 95% probability intervals are wider when considering low credit rating levels (see Figure 2.8). This pattern agrees with the in-sample results reported in Section 2.3, which suggest that the heterogeneity among bonds with a given credit rating increases when considering low rating levels (see Figure 2.3).

2.5.3 Improving Posterior Inference

In general, there are two aspects that can be modified in order to improve the performance of the term structure estimators based on covariates. One aspect is the set of covariates being used to explain the term structure, while the second is the modeling approach used to introduce dependence on the covariates. Regarding the first aspect, there is a rich literature describing methodological tools that can be used for com-

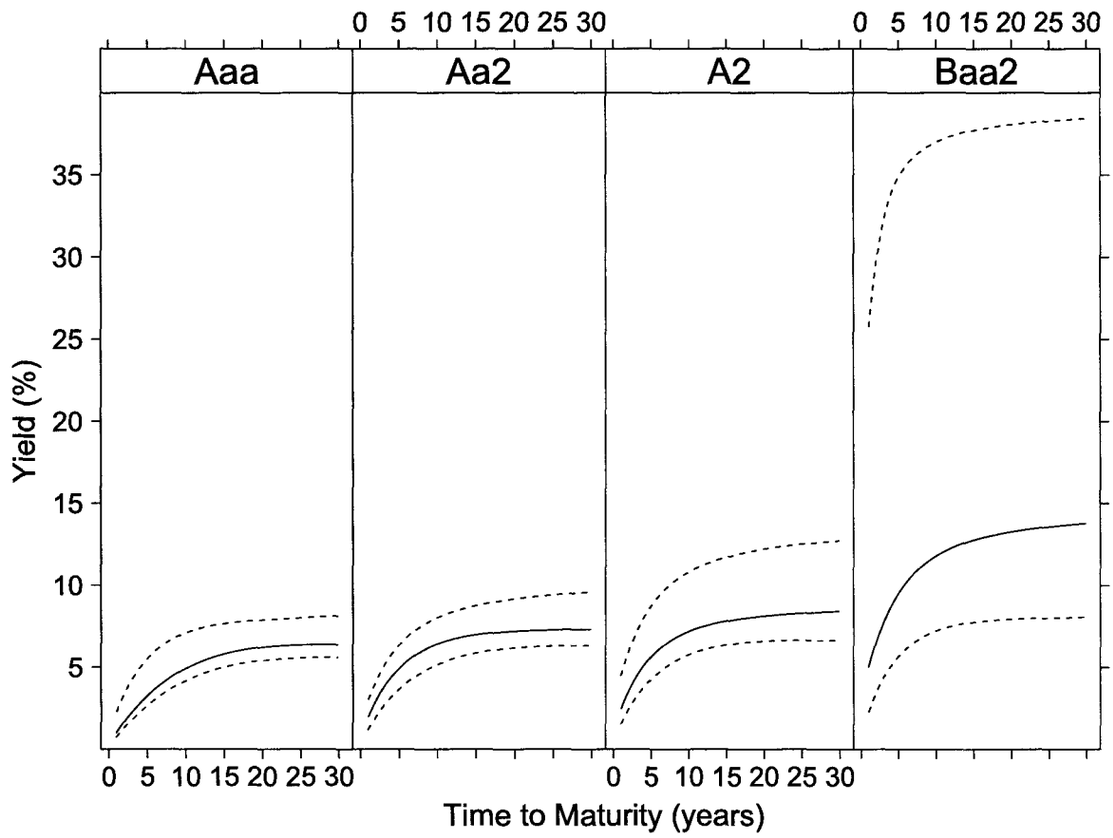


Figure 2.8 : Predictive yield curves for four combination of covariates. Such combinations differ in terms of the Moody's rating level, specifically, the levels Aaa, Aa2, A2 and Baa2 are considered. The other three covariates are the same for the four combinations. The industry group is finance, the credit ratings by Moody's and S&P are the same, and the bond age is greater than one year. The curves denote the median and pointwise 95% probability intervals. The pointwise intervals are wider when considering low credit rating levels. Hence, the heterogeneity among bonds with a given credit rating increases when considering low rating levels.

paring the performance of models corresponding to different sets of covariates (see, for example, Wasserman (2000)). In contrast, limited attention has been directed in the literature towards comparing the performance of different modeling approaches that introduce dependence of covariates in nonparametric Bayesian models.

The next chapter discusses the differences between competing modeling approaches that introduce dependence on covariates. In particular, we explain what features are desirable in such extensions in order to produce better predictions. Although such an analysis is motivated by our interest on predicting the term structure of interest rates, the results described in the next chapter have a broad applicability. That is, they are relevant for any practitioner interested in modeling covariates with nonparametric Bayesian methods.

2.6 Summary and Discussion

We introduced a new framework for the estimation of term structures that produces accurate estimators by jointly estimating multiple term structures. This framework uses a Bayesian population model with a Dirichlet mixture prior that clusters term structures with similar characteristics so that they borrow strength from each other. Each term structure is modeled using a parsimonious representation based on exponential polynomials. Our model can jointly estimate any combination of corporate and/or government term structures. In this chapter, however, we focused on corporate bonds to show the ability of our new framework to produce accurate estimators

based on small samples of bonds, a situation that arises when estimating the term structure for individual corporations.

We used a data set including information for 599 U.S. corporate bonds with transactions executed on June 15, 2009. It is common practice to approximate the term structure of corporate bonds by first grouping the bonds by credit rating and then independently estimating the term structure for each class. We refer to this two-stage method as a single-curve approach. However, we have shown in this chapter that our Bayesian population model can produce more realistic and accurate estimators than the single-curve approach produces. To reach that conclusion we compare the performance of the single-curve estimators with the performance of the estimators produced with our Bayesian population model using two different classification of bonds, respectively. The first classification is based on credit rating and coincides with that used in the single-curve approach, while the second classification is much finer and is based on issuer company. The main results of the two comparisons are summarized below.

When using our Bayesian model to estimate the term structure of bonds grouped by credit rating, the resulting estimators for each credit rating class capture the relationship between credit risk and yield. That is, the lower the credit rating, the higher the yield. In contrast, the single-curve estimators fail to reflect such relationship since they cross each other (see Figure 2.1). Although this empirical evidence already suggests the superior performance of our model, the most important improvement of our

approach over current estimation methods is its ability to estimate term structures by company.

The advantages of grouping the bonds by issuer are noteworthy as evidenced by in-sample and out-of-sample tests. Regarding in-sample goodness of fit, the estimators produced with our method provide a reduction on the average absolute price residual of 78% compared with the single-curve estimators (see Table 2.2 and Figure 2.3). Out-of-sample tests based on cross-validation lead to the same conclusion. Across different partitions, the average of the absolute prediction error is reduced by at least 52% when using our Bayesian model (see Table 2.3 and Figure 2.5).

The empirical evidence shows that our new framework improves the estimation of corporate term structures. However, our model is not restricted to the analysis of corporate bonds. We can consider combinations of corporate and/or government bonds, and consequently, we can easily estimate credit spreads, i.e., the difference between government and corporate yield curves. Computing credit spreads based on the estimators produced with our model are likely to be accurate because of the good performance of our model in identifying the underlying term structure of corporate bonds, along with the fact that there are usually enough government bonds to accurately estimate the risk-free term structure. Finally, our model can also be used to estimate spreads between bonds from different countries. We just need to group the bonds by country, jointly estimate the term structures, and take the difference of the estimated curves by pairs to obtain the spreads.

Regarding the practical implementation of our model, our experience suggest that it does not require excessive tuning. Specifically, we have always been able to initialize the MCMC as well as set the hyper-parameters that are not random in the model by using deterministic rules based on rough estimators of 1) the mean over all the parameters $\boldsymbol{\theta}_i$ and 2) their covariance matrix. A detailed explanation can be found in Appendix A. Finding such estimators is particularly easy when the model has already been used with a data set of the same type as the one of interest but from an earlier date. Specifically, those estimators can be obtained by using the posterior means of $\boldsymbol{\theta}_i$ produced when fitting the model to the "old" data set.

In summary, the estimation model described in this chapter is, to the best of our knowledge, the first model that is able to produce accurate estimators of the term structure when only a handful of bonds are available for each company. Furthermore, it is a flexible model that is not restricted to a specific type of bonds and it can be easily implemented in practice since no excessive tuning of its parameters is required.

Finally, we extended our new framework to introduce dependence on covariates. Such an extension is of interest because it allows us to predict the term structure of interest rates based on a given set of covariates. Although introducing dependence on covariates in nonparametric Bayesian methods has been a very active of research and several modeling approaches have been developed, limited research has examined the relative performance of such methods or improved understanding of which features are suitable in order to produce better results. We focus on those issues in Chapter

3. Although such an analysis is motivated by our interest on predicting the term structure of interest rates, the conclusions are of interest for any practitioner modeling covariates with nonparametric Bayesian methods.

Chapter 3

Modeling Dependence on Covariates

A modeling framework that has become increasingly popular is the use of Bayesian nonparametric methods (Müller et al. 2004). In particular, the Dirichlet process (DP) (Ferguson 1973) is the most popular prior model for an unknown random measure. The popularity of the DP is due to its elegance, simplicity and the existence of computationally efficient Markov chain Monte Carlo (MCMC) posterior simulation algorithms (MacEachern and Müller 1998). In addition, hierarchical models based on random effects distributions with DP priors are able to accommodate outliers, multimodality, and dependence in multivariate, longitudinal, and functional data (Dunson 2009). Examples of nonparametric methods based on the DP include applications in pharmacokinetics (Rosner and Müller 1997), econometrics (Griffin and Steel 2004), spatial modelling (Gelfand et al. 2005), meta-analysis (Burr and Doss 2005), variable selection (Kim et al. 2006), genetics (Xing et al. 2007), density estimation (Dunson et al. 2007; Rodriguez and ter Horst 2008), and survival analysis (De Iorio et al. 2009).

The DP as described by Ferguson (1973) does not incorporate covariates. Hence, an active area of research is extending nonparametric models to allow the unknown distribution to depend on covariates. A popular approach uses as a starting point the Sethuraman (1994) representation of a DP. This representation states that a random

measure G that follows a DP with total mass parameter M and base measure G_0, denoted as $DP(M, G_0)$, can be represented as

$$G = \sum_{h=1}^{\infty} w_h \delta(\theta_h), \quad w_h = v_h \prod_{i=1}^{h-1}(1 - v_i), \quad v_h \sim Beta(1, M), \quad \theta_h \sim G_0, \qquad (3.1)$$

where $\delta(\theta)$ is a point mass at θ. Generalizations of (3.1) achieve the desired dependence of G on covariates by making the weights, w_h, and/or the locations, θ_h, vary with the covariates according to a stochastic process (MacEachern 1999). Such extensions of the Sethuraman representation are probability models on a collection of dependent random probability measures $\{G_x, x \in X\}$, where X is the corresponding covariate space. Generalizations of the Sethuraman representation that introduce covariates via the locations have been applied to the analysis of variance (De Iorio et al. 2004), spatial modeling (Gelfand et al. 2005), time series (Caron et al. 2006), and regression (De Iorio et al. 2009). On the other hand, there are also generalizations based on making the weights covariate-dependent. Griffin and Steel (2006) proposed an order-based dependent Dirichlet process that makes the order of v_h in the stick-breaking construction be a function of the covariates, Dunson and Park (2008) express v_h as a covariate-dependent kernel multiplied by beta weights, and Fuentes-García et al. (2009) models w_h with a nonparametric mixture model that depends on covariates.

Approaches for modeling covariates not based on the Sethuraman representation exist in the literature as well. If the covariates only take a finite number of values, then the product of Dirichlet processes described in Cifarelli and Regazzini (1978) can

be used to introduce dependence on covariates. Specifically, a covariate-dependent regression model is used as the base measure of independent Dirichlet processes at each level of the covariates (Carota and Parmigiani 2002; Griffin and Steel 2004). Another approach for modeling dependence is forming convex combinations of independent DP (Dunson et al. 2007; Müller et al. 2004). Finally, Müller et al. (1996) and Müller and Rosner (1998) include covariates x_i in an augmented response vector (y_i, x_i) and obtain the desired dependence by focusing on the conditional distribution given x_i. We refer to this method as Conditional DP. In this context, y_i denotes a random vector with a DP mixture distribution (see equation (3.2)).

Although modeling dependence on covariates has been a very active area of research, limited research has examined the relative performance of such methods or improved understanding of which features are suitable in order to produce better results. This chapter considers such a comparison, focusing on predictive inference. Different approaches for modeling dependence on covariates can lead to very similar posterior fits and yet produce very different results when used for prediction. In addition, when the predictive density is a mixture, whether or not the weights depend on the covariates plays a major role in determining the quality of the predictions. Such findings are illustrated by comparing the Linear DDP (De Iorio et al. 2009) to the Conditional-DP (Müller et al. 1996); we apply these methods to a simulated data set and to data from a pharmacokinetic meta-analysis. Section 3.1 describes and compares both methods. Implementation and empirical results are reported in

Section 3.2. Finally, conclusion and discussion appear in Section 3.3.

3.1 Modeling Approaches

This section describes the Linear DDP and the Conditional DP, points out their differences regarding the form of the predictive density, and explains how such differences affect the performance of each method.

Both methods introduce continuous covariates into the typical DP mixture (DPM) model given by

$$\boldsymbol{y}_i \overset{iid}{\sim} \int f(\boldsymbol{y}|\boldsymbol{\mu})\, dG(\boldsymbol{\mu}), \qquad G \sim DP(M, G_0), \tag{3.2}$$

where f is a probability density. The DPM model (3.2) is a mixture model with a DP prior on the mixing measure G. In many practical applications, the kernel $f(\boldsymbol{y}|\boldsymbol{\mu})$ is set to be a normal multivariate density with mean $\boldsymbol{\mu}$ and common covariance matrix $\mathbf{S}$.

3.1.1 Linear DDP

The Linear DDP introduced in De Iorio et al. (2009) models the relationship between continuous covariates and the unknown distribution by replacing the random probability measure G in the DPM model (3.2) with a collection of random probability measures indexed by $\boldsymbol{x}$, $\{G_{\boldsymbol{x}}, \boldsymbol{x} \in X\}$, where $\boldsymbol{x} = (x_1, \ldots, x_d)$ denotes a d-dimensional vector of continuous covariates and X is the corresponding covariate space.

De Iorio et al. (2009) set a prior on $\{G_{\boldsymbol{x}}, \boldsymbol{x} \in X\}$ using as a starting point the Dependent DP (DDP), as defined in MacEachern (1999). The DDP specifies

$$G_{\boldsymbol{x}} = \sum_{h=1}^{\infty} w_h \delta(\boldsymbol{\theta}_{\boldsymbol{x}h}), \qquad \text{for any } \boldsymbol{x} \,. \tag{3.3}$$

The point masses $\boldsymbol{\theta}_{\boldsymbol{x}h}$ satisfy the condition that $\theta_h = \{\boldsymbol{\theta}_{\boldsymbol{x}h}, \boldsymbol{x} \in X\}$ are iid realizations of a stochastic process in $\boldsymbol{x}$. The weights, w_h, follow a stick-breaking prior as in Sethuraman (1994), that is, $w_h = v_h \prod_{i=1}^{h-1}(1-v_i)$, $v_h \sim Beta(1, M)$. The DDP model (3.3) implies that, for each x, G_x follows a DP. Specifically, $G_{\boldsymbol{x}} \sim DP(M, G_{0\boldsymbol{x}})$, where the base measure $G_{0\boldsymbol{x}}$ is the marginal distribution at $\boldsymbol{x}$ of the stochastic process on the point masses $\boldsymbol{\theta}_{\boldsymbol{x}h}$.

In general, the DDP model induces dependence of the random measures $G_{\boldsymbol{x}}$ by assuming that the sample paths θ_h are dependent across $\boldsymbol{x}$. De Iorio et al. (2009), in particular, impose a linear model on $\boldsymbol{\theta}_{\boldsymbol{x}h}$ given by

$$\boldsymbol{\theta}_{\boldsymbol{x}h} = \boldsymbol{m}_h + \sum_{i=1}^{d} \boldsymbol{\beta}_{ih} x_i, \tag{3.4}$$

where d is the number of continuous covariates, $\boldsymbol{m}_h \overset{iid}{\sim} p^0_{\boldsymbol{m}}$ and $\boldsymbol{\beta}_{ih} \overset{iid}{\sim} p^0_{\boldsymbol{\beta}_i}$, $i = 1 \ldots d$. It follows that the base measure, $G_{0\boldsymbol{x}}$, is given by the convolution of $p^0_{\boldsymbol{m}}$ and $p^0_{\boldsymbol{\beta}_i}$, $i = 1 \ldots d$. The distributional assumptions on $\boldsymbol{\theta}_{\boldsymbol{x}h}$ proposed by De Iorio et al. (2009) imply that the random measures $G_{0\boldsymbol{x}}$ share the common main effect given by $\boldsymbol{m}_h$. In addition, for each i, $\boldsymbol{\beta}_i$ represents a slope coefficient as in a standard linear model.

The prior given in (3.3) with the linear model on the locations as in (3.4) is the Linear DDP (De Iorio et al. 2009). When such a prior is used to introduce continuous

covariates into the DPM model (3.2) with a normal kernel, it leads to models given by

$$\begin{aligned} (\boldsymbol{y}_i|\boldsymbol{x}_i = \boldsymbol{x}) &\overset{iid}{\sim} \int N(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{S})\, dG_{\boldsymbol{x}}(\boldsymbol{\mu}), \\ \{G_{\boldsymbol{x}}, \boldsymbol{x} \in X\} &\sim \text{Linear DDP}(M, G_{0X}), \end{aligned} \tag{3.5}$$

where G_{0X} denotes the set of distributions for the main effect and slope coefficients in (3.4). The model is completed with appropriate hyperpriors for S, M, and G_{0X}.

It is possible to rewrite model (3.5) in terms of a mixture of linear models. Specifically, we define the random matrix $\tilde{\boldsymbol{\Gamma}}_h = [\boldsymbol{m}_h, \boldsymbol{\beta}_{1h}, \ldots, \boldsymbol{\beta}_{ph}]$ and design vectors $\boldsymbol{d}_i = (1, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, such that $\boldsymbol{\theta}_{\boldsymbol{x}_i h} = \tilde{\boldsymbol{\Gamma}}_h \boldsymbol{d}_i$, where $\boldsymbol{x}_i$ denote the continuous covariate vector of subject i. Model (3.5) can then be rewritten as

$$\begin{aligned} (\boldsymbol{y}_i|\boldsymbol{x}_i) &\overset{iid}{\sim} \int N\left(\boldsymbol{y}; \tilde{\boldsymbol{\Gamma}} \boldsymbol{d}_i, \mathbf{S}\right) dG(\tilde{\boldsymbol{\Gamma}}), \\ G &\sim DP(M, G_0), \end{aligned} \tag{3.6}$$

with base measure $G_0 = \left(p^0_{\boldsymbol{m}}, p^0_{\boldsymbol{\beta}_1}, \ldots, p^0_{\boldsymbol{\beta}_p}\right)$. The reformulation (3.6) is convenient because it has the form of a DPM model (3.2). Hence, it is possible to carry out posterior inference in the Linear DDP model by using well-known MCMC algorithms designed for DPM models, such as those described in MacEachern and Müller (1998).

For later reference we replace the mixture in (3.6) by an additional level in the hierarchical model

$$\begin{aligned} (\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Gamma}_i, S) &\overset{iid}{\sim} N(\boldsymbol{y}; \boldsymbol{\Gamma}_i \boldsymbol{d}_i, \mathbf{S}), \\ \boldsymbol{\Gamma}_i \sim G, \quad &\text{and} \quad G \sim DP(M, G_0). \end{aligned} \tag{3.7}$$

Since a DP is almost surely discrete, there is a positive probability for ties among

the $\mathbf{\Gamma}_i$. Let $\{\mathbf{\Gamma}_1^\star \ldots \mathbf{\Gamma}_k^\star\}$, $k \leq n$, be the set of unique values in $\{\mathbf{\Gamma}_1 \ldots \mathbf{\Gamma}_n\}$, and let n_j denote the number of $\mathbf{\Gamma}_i$ equal to $\mathbf{\Gamma}_j^\star$.

We conclude the description of the Linear DDP by deriving the predictive density of model (3.5). Let Θ denote the set of all model parameters and let $\Theta^{(t)}$, $t = 1, \ldots, N$ denote a posterior Monte Carlo sample. We will generically use the superscript (t) to indicate elements of $\Theta^{(t)}$. For a new subject with vector of covariates $\boldsymbol{x}_{n+1}$, the predictive density can be approximated as follows:

$$\begin{aligned} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y\right) &= E\left[p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y, \Theta\right)|Y\right] \\ &\approx \frac{1}{N}\sum_{t=1}^{N} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y, \Theta^{(t)}\right) \\ &= \frac{1}{N}\sum_{t=1}^{N} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right), \end{aligned}$$

where Y denotes the current data. The specific expressions for the probabilities in the last average are easily obtained from (3.7).

$$p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right) = \sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)}\, p_j\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right) \tag{3.8}$$

with

$$\alpha_j^{(t)} \propto \begin{cases} n_j^{(t)} & j = 1 \ldots k^{(t)}, \\ M^{(t)} & j = k^{(t)} + 1, \end{cases}$$

where $\sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)} = 1$, and

$$p_j\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right) = \begin{cases} N\left(\boldsymbol{y}_{n+1}; \mathbf{\Gamma}_j^{\star(t)}\boldsymbol{d}_{n+1}, \boldsymbol{S}^{(t)}\right) & j = 1 \ldots k^{(t)}, \\ \int N\left(\boldsymbol{y}_{n+1}; \mathbf{\Gamma}\boldsymbol{d}_{n+1}, \boldsymbol{S}^{(t)}\right) dG_0(\mathbf{\Gamma}) & j = k^{(t)} + 1. \end{cases}$$

3.1.2 Conditional DP

The Conditional DP approach introduces regression into the DPM model (3.2) by including the continuous covariates in an augmented response vector $\tilde{\boldsymbol{y}} = (\boldsymbol{y}, \boldsymbol{x})$ in the nonparametric model. Specifically, Müller and Rosner (1998) make the unknown distribution $p(\boldsymbol{y})$ depend on covariates $\boldsymbol{x}$ by defining a DP mixture model for the joint model $p(\boldsymbol{y}, \boldsymbol{x})$. Considering a normal kernel, the Conditional DP modifies the DPM model as follows:

$$(\boldsymbol{y}_i, \boldsymbol{x}_i) \overset{iid}{\sim} \int N\left((\boldsymbol{y}, \boldsymbol{x}); \boldsymbol{\mu}, \mathbf{S}\right) dG(\boldsymbol{\mu}), \qquad G \sim DP(M, G_0), \tag{3.9}$$

which is equivalent to the hierarchical model

$$\begin{aligned} (\boldsymbol{y}_i, \boldsymbol{x}_i) &\overset{iid}{\sim} N\left((\boldsymbol{y}, \boldsymbol{x}); \boldsymbol{\mu}_i, \mathbf{S}\right), \\ \boldsymbol{\mu}_i \sim G, \quad &\text{and} \quad G \sim DP(M, G_0). \end{aligned} \tag{3.10}$$

We refer to this modeling approach as Conditional DP, because the implied conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ formalizes the desired regression on $\boldsymbol{x}$. Particularly, as explained in Müller and Rosner (1998), the mixture of normals $\int N\left((\boldsymbol{y}, \boldsymbol{x}); \boldsymbol{\mu}, \mathbf{S}\right) dG(\boldsymbol{\mu})$ implies a locally weighted mixture of normal linear regressions for $E[\boldsymbol{y}_i|\boldsymbol{x}_i]$. Implementation of posterior inference for the Conditional DP model (3.9) is straightforward, because it has the form of a DPM model. Hence, the MCMC algorithms described in MacEachern and Müller (1998) can be used.

The predictive density for a new $(n+1)$-th subject, conditional on covariates x_{n+1},

is estimated using

$$p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y\right) \approx \frac{1}{N}\sum_{t=1}^{N} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right). \tag{3.11}$$

The conditional density in the last average is derived from the density $p\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$ corresponding to the DPM model (3.10) as follows. Let $\{\boldsymbol{\mu}_1^\star \ldots \boldsymbol{\mu}_k^\star\}$, $k \leq n$, be the set of distinct vectors from the set $\{\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n\}$, with n_j the number of $\boldsymbol{\mu}_i$ equal to $\boldsymbol{\mu}_j^\star$. The predictive density for $\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}\right)$ is

$$p\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right) = \sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)}\, p_j\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right) \tag{3.12}$$

with

$$\alpha_j^{(t)} \propto \begin{cases} n_j^{(t)} & j = 1 \ldots k^{(t)}, \\ M^{(t)} & j = k^{(t)} + 1, \end{cases}$$

where $\sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)} = 1$, and

$$p_j\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right) = \begin{cases} N\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}; \boldsymbol{\mu}_j^{\star(t)}, \boldsymbol{S}^{(t)}\right) & j = 1 \ldots k^{(t)}, \\ \int N\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}; \boldsymbol{\mu}, \boldsymbol{S}^{(t)}\right) dG_0(\boldsymbol{\mu}) & j = k^{(t)} + 1. \end{cases}$$

It follows that the conditional density needed in (3.11) is given by

$$p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right) = \sum_{j=1}^{k^{(t)}+1} \left(\frac{\alpha_j^{(t)}\, p_j\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)}{\sum_{\ell=1}^{k^{(t)}+1} \alpha_\ell^{(t)}\, p_\ell\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)} \right) p_j\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right), \tag{3.13}$$

where $p_j\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$ is obtained by integrating out $\boldsymbol{y}_{n+1}$ from the joint distribution $p_j\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$ in the mixture (3.12).

3.1.3 Comparing Approaches

Both extensions of the DPM model, the Linear DDP and the Conditional DP, estimate the predictive density by averaging the mixture distribution $p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right)$ with respect to a Monte Carlo sample from the posterior distribution. We claim, however, that the specific factors determining the weights in such mixtures can affect the resulting predictive inference.

As shown in (3.8), the weights in the mixture density $p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right)$ corresponding to the Linear DDP are solely determined by $n_j^{(t)}$ and $M^{(t)}$. That is, the relative importance of each component in the mixture is mainly determined by the size of the "clusters" induced by the unique values of $\Gamma_i^{(t)}$. This feature is an important limitation of the Linear DDP. It implies that, for each t, the weights of the mixture remain the same for any new subject rather than change as a function of the specific covariates x_{n+1}. In other words, there is no built-in mechanism in the predictive density to favor those components in the mixture that are more likely to provide a better fit to the specific characteristics of the new subject.

Unlike the formulas in the Linear DDP, the weights in the mixture $p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right)$ corresponding to the Conditional DP are a function of the covariates via the marginal density $p_j\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$, as shown in (3.13). It follows that components in the mixture with a high marginal density on $\boldsymbol{x}_{n+1}$ will tend to have higher weights. Hence, the mixture is adjusted to reflect the specific characteristics of a given new subject. The inclusion of covariates in the weights increases the chance of using a regression struc-

ture suitable for the given subject.

In summary, the differences mentioned above suggest that, in terms of predictive inference, the Conditional DP should outperform the Linear DDP. We present empirical evidence corroborating such a claim in the next section.

Finally, we comment on a weakness of the Conditional DP that somewhat offsets the discussed features. The sampling model in (3.8) can be factored as $p(y_i|x_i) * p(x_i)$, highlighting the fact that the likelihood includes an additional factor for the covariates x_i. This is technically inappropriate when the covariates are chosen and fixed by design.

3.2 Empirical Implementation

This section provides evidence to illustrate the superior performance of the Conditional DP over the Linear DDP for predictive inference. Two data sets are considered; the first one is a simulated data set, while the second derives from a population pharmacokinetic study.

3.2.1 Simulation Example

The simulated data set corresponds to a multiple regression model with two dependent variables and a single predictor variable. Specifically, the data are generated from a model similar to (3.6) as follows. Let $\boldsymbol{\Gamma} = [\boldsymbol{m}, \boldsymbol{\beta}]$ be a 2×2 matrix where $\boldsymbol{m}$ is a vector of constants and $\boldsymbol{\beta}$ is a vector of slope coefficients. Our simulated data set

includes a sample of bivariate observations $\boldsymbol{y}_i = (y_{i1}, y_{i2})^{\mathrm{T}}$, $i = 1, \ldots, 100$, generated from the bivariate distribution $N(\boldsymbol{\Gamma}_i \boldsymbol{d}_i s\boldsymbol{I})$, where $\boldsymbol{I}$ is the 2×2 identity matrix, $s = 0.10$, and $\boldsymbol{d}_i = (1, x_i)^{\mathrm{T}}$ is a design vector including the subject-specific covariate x_i. We introduce two underlying regression structures into the simulated data set by randomly setting $\boldsymbol{\Gamma}_i$ equal to one of the following:

$$\begin{cases} \boldsymbol{\Lambda}_1 = \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix}, & \text{w.p. } 1/2 \\ \\ \boldsymbol{\Lambda}_2 = \begin{bmatrix} 0 & 0 \\ 1 & 3 \end{bmatrix}, & \text{w.p. } 1/2 \end{cases} \tag{3.14}$$

Finally, each x_i is generated from a uniform distribution. If $\Gamma_i = \Lambda_1$ then $x_i \sim \mathrm{U}(0, 1)$, otherwise $x_i \sim \mathrm{U}(-1, 0)$. Hence, the value of the covariate x_i provides information about the specific underlying regression structure. As shown later, the ability to incorporate such information is a crucial difference between the Linear DDP and the Conditional DP.

The resulting simulated data set reflects two regression structures given by the product $\boldsymbol{\Gamma}_i \boldsymbol{d}_i$; one describes the horizontal line segment $(0, 1)^{\mathrm{T}} + x(3, 0)^{\mathrm{T}}$, $x \in (0, 1)$, while the other follows the vertical line segment $(0, 1)^{\mathrm{T}} + x(0, 3)^{\mathrm{T}}$, $x \in (-1, 0)$ (see Figure 3.1). The simulated data are generated from a model that resembles (3.6). Thus, the data set matches the modeling assumptions of the Linear DDP. Such a feature was chosen in order to rule out the characteristics of the simulated data as an

explanation for the poor performance of the Linear DDP.

We used posterior MCMC simulation to generate posterior Monte Carlo samples under Linear DDP (3.6) and the Conditional DP (3.9), respectively. In both cases the uncertainty on the common matrix $\boldsymbol{S}$ is modeled by adopting the conjugate inverse Wishart prior $\boldsymbol{S}^{-1} \sim Wishart(r, (r\boldsymbol{R})^{-1})$ with r degrees of freedom and mean $r(r\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}$. In addition, the base measure is assumed to be multivariate normal, $G_0 \sim N(\boldsymbol{b}, \boldsymbol{B})$, and M is given a gamma distribution, $M \sim Ga(a_m, b_m)$. In both cases, we considered 10,000 iterations of the MCMC algorithm; convergence of such algorithms was reached after 2000 iterations. In general, the posterior fit under the Linear DDP is slightly better than the one obtained with the Conditional DP. For example, we estimated the first, second, and third quartiles, (Q1,Q2,Q3), of the distribution of the Euclidean distance between the posterior mean of $\boldsymbol{y}_i$ and its sample value. Those statistics were equal to $(0.07, 0.12, 0.18)$ for the Linear DDP, while the Conditional DP showed somewhat greater values, $(0.09, 0.16, 0.32)$.

Substituting the posterior Monte Carlo samples in (3.8) and (3.13), we evaluated the posterior predictive density corresponding to a new subject with covariate $x \in (-1, 1)\backslash\{0\}$. The results for $x = 0.5$ and $x = -0.5$ highlight the shortcoming of the Linear DDP. It wrongly assigns positive probability to regions in the plane without any sample points (see Figure 3.1). Such results can be explained as follows. Since the Linear DDP correctly recognizes the two regression structures, the mixture in (3.8) is dominated by two components, each one of them corresponding to one of the

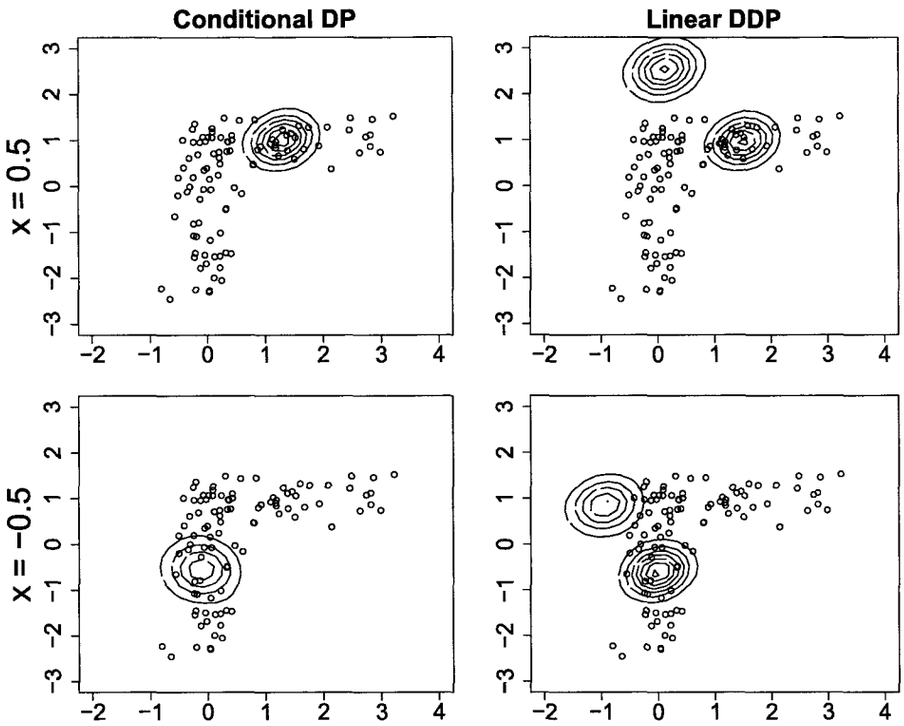


Figure 3.1 : Contours of the predictive density. The elements of the simulated data set are marked with dots. The plots are organized by modeling approach (columns) and by covariate level used to perform predictive inference (rows). The estimators produced by the Linear DDP are not acceptable because they wrongly introduce an extra mode to the predictive density. Such results reflect the fact that the Linear DDP selects a regression structure based solely on the cluster size, which by design is approximately half for each one of the two underlying regression structures, rather than using the covariate level of the new subject.

regression structures. It follows that for any given covariate x, the mixture in (3.8) gives positive probability to two regions, one for each regression structure. In contrast, the Conditional DP incorporates the values of the covariates into the weights of the mixture (3.13). Therefore, the Conditional DP is able to identify those regression structures that are more likely, given the value of x.

3.2.2 Pharmacokinetics Example

We consider a data set that includes clinical covariates and longitudinal data consisting of drug concentration measurements. The measurements came from patients receiving the anti-cancer drug topotecan in several different studies. We fit a Bayesian population pharmacokinetic (PK) model to estimate and predict plasma concentration-time curves. This section compares the performance of the Linear DDP and Conditional DP when introducing dependence on patient covariates into the PK model. The inclusion of covariates enable us to predict individual concentration-time curves appropriate for an individual patient instead of the group average. Such estimated curves provide valuable information that can be used, for example, to design dose individualization schemes for future patients who are starting topotecan treatment.

Before describing the Bayesian PK model, here are some details about the the data set. The population consisted of 138 children enrolled in seven clinical studies (see Table 3.1). These data have previously been used by Schaiquevich et al. (2007) to characterize the population pharmacokinetics of topotecan lactone in children with cancer and to identify covariates related to topotecan disposition. The data include concentration-time measurements corresponding to several treatment occasions for each patient. We use only those corresponding to the first treatment in our sample. This is because we are interested in making inference (and prediction) for patients with no previous topotecan treatment. In all cases, topotecan was administered via IVAC Controller (IVAC corp.) with duration of infusion of 30 minutes. See Schaiquevich

et al. (2007) and references therein for more details regarding the eligibility criteria for the clinical studies, drugs administration, blood collection, and patient demographics.

The structure of a Bayesian population PK model is as follows. Let y_{ij} denote the j-th measurement for the ith patient, and $\boldsymbol{\theta}_i$ the vector of random effects of patient i. The vector $\boldsymbol{x}_i$ represents the patient-specific covariates. The probability model for the PK data is given by

$$p(y_{ij}|\boldsymbol{\theta}_i), \qquad p(\boldsymbol{\theta}_i|\phi), \qquad p(\phi). \tag{3.15}$$

Here, $p(y_{ij}|\boldsymbol{\theta}_i)$ is a parametric, and typically non-linear, regression model for the concentration-time curve, $p(\boldsymbol{\theta}_i|\phi)$ is the prior for $\boldsymbol{\theta}_i$, and, finally, $p(\phi)$ denotes the probability model of the hyperparameters. Bayesian models similar to (3.15) have been considered in Zeger and Karim (1991) for generalized linear mixed models and in Wakefield (1994) using a multivariate normal population distribution as population model.

Both methods, the Linear DDP and the Conditional DP, can be used to introduce covariates into the PK model (3.15) via the prior distribution on $\boldsymbol{\theta}_i$. The resulting prior has the form (3.5) or (3.9), respectively, where $\boldsymbol{y}_i$ is replaced with $\boldsymbol{\theta}_i$. That is, using the Linear DDP model, the prior on $(\boldsymbol{\theta}_i|\boldsymbol{x}_i)$ is a mixture of normals with a mixing measure given by a family of random measures indexed by $\boldsymbol{x}$. If the Conditional DP is used, the vector of parameters is augmented to include the covariates. Hence, the prior $p(\boldsymbol{\theta}_i, \boldsymbol{x}_i)$ is a mixture of normals

$$p\left(\boldsymbol{\theta}_i, \boldsymbol{x}_i\right) = \int N\left(\left(\boldsymbol{\theta}_i, \boldsymbol{x}_i\right); \boldsymbol{\mu}, \mathbf{S}\right) dG(\boldsymbol{\mu}),$$

Table 3.1 : Characteristics of the clinical trials from which PK data was obtained

Trial no.	No. of patients	Trial type	Topotecan dosage
1	15	Phase I recurrent solid tumors	Target AUC = 120-180 ng · h/mL
2	21	Phase I recurrent acute leukemia	Fixed dosage = 2.4 mg/m^2
3	28	Phase I recurrent solid tumors	0.8 and 1.1 mb/m^2
4	10	Phase II newly diagnosed medulloblastoma	Target AUC = 120-160 ng · h/mL
5	22	Phase I recurrent solid tumors	1.4, 1.7, 2.0, and 2.4 mg/m^2
6	30	Phase II newly diagnosed high-risk medulloblastoma	Target AUC = 80-120 ng · h/mL
7	12	Phase II recurrent Wilms tumor	Target AUC = 70-90 ng · h/mL

with a DP prior on the mixing measure G. The model achieves the desired nonlinear, semi-parametric regression of the parameters on the covariates via the conditional distribution

$$p(\boldsymbol{\theta}_i|\boldsymbol{x}_i) = \frac{p(\boldsymbol{\theta}_i, \boldsymbol{x}_i)}{\int p(\boldsymbol{\theta}_i, \boldsymbol{x}_i) d\boldsymbol{\theta}_i} \propto p(\boldsymbol{\theta}_i, \boldsymbol{x}_i).$$

Both modeling approaches assign a flexible nonparametric prior distribution to the population parameters. Hence, they both are able to accommodate heterogeneity in the patient population, such as outliers, over-dispersion, and multimodality.

Since we want to emphasize the differences between the Linear DDP and the Conditional DP, similar criteria are used to specify $p(y_{ij}|\boldsymbol{\theta}_i)$ and $p(\phi)$, regardless of the prior on $\boldsymbol{\theta}_i$ being considered.

The model for the concentration-time curve, $p(y_{ij}|\boldsymbol{\theta}_i)$, is determined by the non-linear regression

$$\log(y_{ij}) = \log(f(\boldsymbol{\theta}_i, \tau_{ij})) + \epsilon_{ij},$$

where y_{ij} is the jth concentration measurement for the ith patient at time τ_{ij}, $\epsilon_{ij} \sim N(0, \lambda^{-1})$ is the noise term with precision λ, and the function f is a two-compartment model with constant rate intravenous infusion (Wagner 1968) that describes the concentration at time τ as

$$\begin{aligned} f(\boldsymbol{\theta}, \tau) = \frac{D/\gamma}{V_1 K_2} \Bigg\{ 1 - \Bigg[\left(\frac{K_2 - \alpha}{\beta - \alpha}\right) e^{-\beta\tau + \beta(\tau-\gamma)\mathbf{1}_{(\tau \geq \gamma)}} \\ - \left(\frac{K_2 - \beta}{\beta - \alpha}\right) e^{-\alpha\tau + \alpha(\tau-\gamma)\mathbf{1}_{(\tau \geq \gamma)}} \Bigg] \Bigg\} \times \\ \left[\left(\frac{K_2 - \alpha}{\beta - \alpha}\right) e^{-\beta(\tau-\gamma)} - \left(\frac{K_2 - \beta}{\beta - \alpha}\right) e^{-\alpha(\tau-\gamma)} \right]^{\mathbf{1}_{(\tau \geq \gamma)}} \end{aligned} \tag{3.16}$$

with

$$\begin{aligned}\alpha &= \frac{1}{2}\left[(K_1+K_2+K_{-1})+\sqrt{(K_1+K_2+K_{-1})^2-4K_{-1}K_2}\right],\\ \beta &= \frac{1}{2}\left[(K_1+K_2+K_{-1})-\sqrt{(K_1+K_2+K_{-1})^2-4K_{-1}K_2}\right].\end{aligned}$$

Here, the four parameters $\{V_1, K_2, K_1, K_{-1}\}$ are all positive, γ is the duration of infusion, D is the dose, and $\mathbf{1}_{(\cdot)}$ denotes an indicator function. Finally, we parameterize (3.16) with

$$\boldsymbol{\theta} = (log(V_1), log(K_2), log(K_1), log(K_{-1}))^{\mathrm{T}}.$$

Such a parameterization guarantees that the subject-specific parameters can take any value, and thus the use of a mixture of normals as a prior is appropriate.

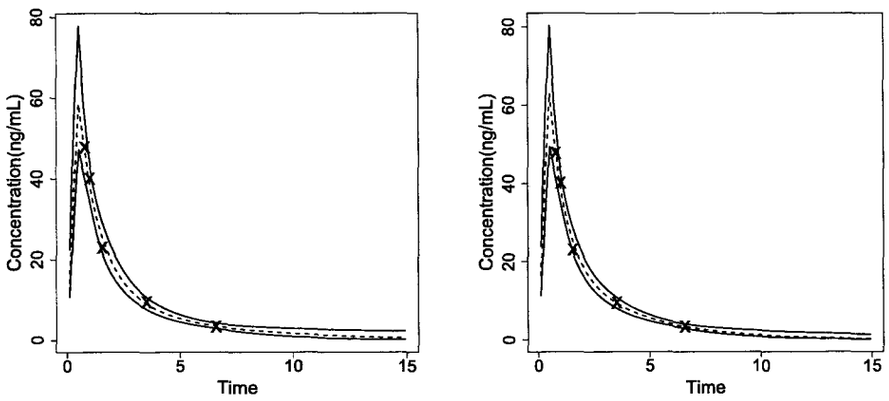
Finally, we introduce distributional assumptions for the hyperparameters, ϕ, that allow the implementation of a MCMC algorithm that is computationally efficient. In the discussion that follows, it is assumed that the Linear DDP is written in the equivalent form (3.6). Regarding the parameters of the DP, the total mass parameter M is given a gamma prior $Ga(a_m, b_m)$, while the base measure G_0 follows a multivariate normal distribution $N(\boldsymbol{b}, \boldsymbol{B})$. The moments of G_0 are assumed random with hyperpriors $\boldsymbol{b} \sim N(\boldsymbol{b_0}, \boldsymbol{B_0})$ and $\boldsymbol{B}^{-1} \sim Wishart\,(w, (w\boldsymbol{W})^{-1})$, where w is the degrees of freedom and $\boldsymbol{W}^{-1}$ is the mean of $\boldsymbol{B}^{-1}$. Finally, a Wishart prior is also used for $\boldsymbol{S}$, $\boldsymbol{S}^{-1} \sim Wishart(r, (r\boldsymbol{R})^{-1})$, and λ is given a gamma prior $G(a_\lambda, b_\lambda)$.

Unlike $p(y_{ij}|\boldsymbol{\theta}_i)$, the characteristics of $p(\phi)$ change in accordance with the prior being used for $\boldsymbol{\theta}$. Specifically, the dimension of G_0 changes as follows. Let p denote the number of model parameters; that is, p equals the dimension of $\boldsymbol{\theta}$. Let d be the

number of covariates being modeled. For the Conditional DP, G_0 is a distribution on a vector with dimension $p + d$, while under the Linear DDP, G_0 is a distribution on the columns of the matrix $\mathbf{\Gamma}$. In the latter case, it is useful to think of G_0 as the distribution on the vector with dimension $p\ (d + 1)$ which results from stacking the columns of $\mathbf{\Gamma}$ one on top of the other.

Putting together the assumptions described above for each component in (3.15), it follows that two population PK models have been completely specified, each one implementing a different prior on $\boldsymbol{\theta}$. For brevity, we will refer to those PK models by using only the name of the modeling approach used for the prior on $\boldsymbol{\theta}$, that is, Conditional DP or Linear DDP. The analysis that follows includes the covariates age, body surface area (BSA), and glomerular filtration rate (GFR). Such covariates were chosen because, as shown in Schaiquevich et al. (2007), they are significant for explaining topotecan disposition. The observed measurements for each covariate were centered at zero.

Implementation of both PK models requires a MCMC scheme to sample from the corresponding posterior distribution. Such sampling schemes can be efficiently implemented, since both the kernel of the mixture and the base measure G_0 are normally distributed (MacEachern and Müller 1998). Since the full conditional of all the parameters, with the exception of $\boldsymbol{\theta}$, have a closed form, they can be updated via Gibbs sampling. For $\boldsymbol{\theta}$, the non-linearity in (3.16) implies that its full conditional is not a known distribution. Therefore, $\boldsymbol{\theta}$ is updated using the adaptive Metropolis Hasting



(a) Conditional DP (b) Linear DDP

Figure 3.2 : Posterior concentration-time curves by method. The dashed lines denote the posterior mean concentration curves, the solid lines refer to the 95% pointwise probability intervals, and the "x"s denote the observed concentration-time combinations. The estimated curves show that the posterior fits produced by both methods are practically the same.

algorithm introduced by Haario et al. (2001). For each model, the corresponding MCMC sampling scheme is used to draw a Monte Carlo sample of size 30,000 from the posterior distribution, after a burn-in period of 20,000 iterations.

Despite the differences between the Linear DDP and the Conditional DP, they lead to very similar posterior fits on the concentration-time curves per child. Specifically, posterior mean estimates of each curve, along with pointwise 95% probability intervals, are practically indistinguishable (see Figure 3.2). Although the curves correspond to a single patient in the data set, similar results are found for the 138 children in the study population.

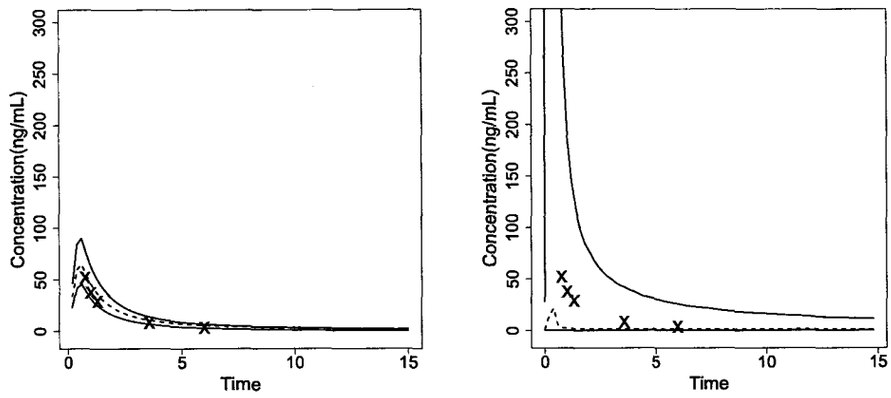
It is in terms of prediction, however, that the Linear DDP and the Conditional DP show different performance. In particular, we compare the predictive distribution for

the concentration-time curves of the 15 patients belonging to study 1. We obtained a Monte Carlo sample from those distributions by first fitting each PK model to the other patients in the data set (123 children in studies 2 - 7), then generating a Monte Carlo sample from the predictive distribution of $\boldsymbol{\theta}$ for each patient belonging to study 1 (using formulas (3.8) and (3.13), along with the patient-specific covariates), and, finally, using each of those samples to evaluate (3.16).

Comparison of the results by method shows that only the Conditional DP is able to produce sensible estimators. As seen in Figure 3.3, the Linear DDP has greater predictive uncertainty than the Conditional DP and led to unrealistic predicted concentration-time profiles. When using the Linear DDP, the estimated predictive distribution provides no information about the real concentration-time curves. Although the curves shown correspond to a single patient, similar results were found for all the patients in study 1.

3.3 Summary and Discussion

In this chapter we have focused on extensions of nonparametric Bayesian models that introduce dependence on covariates. We have shown that good posterior fits with those extensions do not necessarily translate to good prediction. In addition, we have shown that, when the predictive density of such extensions is estimated by averaging mixture distributions, better predictions are produced when the weights in that mixture depend on the covariates. The arguments above have been illustrated



(a) Conditional DP (b) Linear DDP

Figure 3.3 : Predictive concentration-time curves by method. The curves denote the pointwise quartiles Q1 $\leq$ Q2 $\leq$ Q3, and the "x"s mark the observed concentration-time combinations.

by comparing the Linear DDP (De Iorio et al. 2009) and the Conditional DP (Müller and Rosner 1998).

When using the Linear DDP, the weights in the predictive density are solely determined by the size of the clusters induced by the DP, with each cluster representing a different regression structure. Such a feature is a drawback, because it leads to the Linear DDP being unable to identify those regression structures that are more likely for a new subject, given that subject's and the other observed covariates. Combining the covariates with inadequate regression structures leads to unrealistic estimators. Evidence of such behavior has been shown in this chapter. In the simulation example, the Linear DDP wrongly introduced an additional mode to the estimated predictive density, while in the pharmacokinetic example, it predicted unrealistic concentration curves.

In contrast, the predictive density of the Conditional DP is a mixture with weights that depend on the covariates of the new subject. This results in an agreement between the covariates and the regression structures. Such an agreement leads to considerable improvement in the predictive inference when compared to the Linear DDP. In spite of the clear differences in terms of prediction, both methods produce identical results when used for posterior inference.

Although both the Linear DP and Conditinal DDP introduce dependence on continuous covariates, it is also desirable when modeling categorical and discrete covariates to be able to identify the components in the predictive mixture that produce the best fit. Hence, in those cases it is also advantageous to have predictive densities with covariate-dependent weights. However, under some modeling approaches for introducing dependence on finite covariates (numeric or categorical), such as the ANOVA DDP (De Iorio et al. 2004) or the Spatial DDP (Gelfand et al. 2005), the limitations from not having covariate-dependent weights may not be as noticeable as those shown by the Linear DDP. This happens because, in those models, the mechanism through which the covariates determine the form of the components in the mixture can vary, if needed, depending on the specific values of the covariates; this is not achieved by the Linear DDP because, in each component, the corresponding regression structure remains the same regardless of the specific values of the continuous covariates. In particular, the ANOVA DDP models dependence on categorical covariates in an ANOVA fashion. Assuming, for simplicity, that there is only one categorical covariate,

the ANOVA DDP implies that the mean of each component in the predictive mixture distribution is equal to a mean effect plus an offset vector. Since such an offset vector varies according to the value of the covariate, the means of the mixture components automatically account for component-specific associations with the discrete covariate level. Hence, it is possible to generate reasonable inferences, even though the weights in the predictive density are only based on the size of the clusters induced by the DP.

Finally, the results in this chapter provide insight regarding which strategies for introducing dependence on covariates can lead to better posterior inferences. Specifically, it provides evidence in favor of extensions of the Sethuraman representation that make the weights vary with the covariates, because such extensions result in nonparametric Bayesian models with a prediction rule based on a mixture with covariate-dependent weights. Example of those extensions include the order-based DDP (Griffin and Steel 2006) and the kernel stick-breaking process (Dunson and Park 2008).

Chapter 4

Conclusion

Estimating the term structure of corporate bonds generally consists of grouping bonds based on their credit rating level and within each rating class fitting the Nelson-Siegel or some similar functional form to the discount curve using non-linear least squares. This estimation method relies on the assumption, that once the bonds are grouped by rating class they represent a homogeneous group of bonds that can be characterized by one overall model for the term structure. In general, this assumption does not hold. Characterizing groups of bonds by other features than their rating results in homogenous groups but also yields small sample sizes within each grouping. For example, if we group bonds by the company that issued the bond, many companies issue only one bond within a given day of trading.

Utilizing the framework of a Bayesian hierarchical model, we develop a different modeling and estimation strategy that is not limited by the small sample size within each grouping. Among other features, the Bayesian hierarchical model allows the parameters within each small group to be modeled by a different mean thereby greatly reducing the bias in the overall estimation of the term structure. Estimation is accomplished through internal and nonparametric clustering of bonds that exhibit similar structure, thereby allowing the estimation algorithm to borrow strength when appro-

priate across all bonds to identify the structure at the small group level. Specifically, the common population distribution on the parameters is modeled with a Dirichlet process mixture. Applying our newly developed methodology to estimating bonds based on the company that issued the bond, we see a 75% reduction in the mean squared error of out of sample price estimation compared to the traditional approach of grouping by rating class.

In addition to model fitting, we also address this issue of predictions from the resulting model. A challenge for hierarchical nonparametric Bayesian methods is correctly accounting for the clusters with small probabilities when proceeding with Monte Carlo samples to obtain the predictive posterior distribution. We argue that better predictions are produced when the corresponding predictive mixture density has covariate-dependent weights. This dependence on covariates allows for identification of the best set of covariates available for prediction. This argument is illustrated with an application in a biological setting, demonstrating the universal features of the statistical methodology.

In the course of our work, we have identified the following areas of future research. First, we will extend our covariate based predictive modeling to identifying the best set of covariates that can be used to predict the term structure. Secondly, the framework introduced in this dissertation is flexible and can be used with different type of bonds and different groupings of bonds; we can jointly fit term structures of different countries, or, as shown in this document, fit term structures of different

companies. However, such flexibility implies that we are not taking advantage of the specific characteristics of the bonds under consideration. Therefore, an area of future research is to modify our estimation framework to match the specific characteristics of the estimation problem in turn. For example, the term structure of corporate bonds can be expressed as the sum of the government term structure and a credit spread. Modeling credit spreads instead of the corporate term structure could lead to better results since credit spreads can be represented with functional forms based on fewer parameters than the whole term structure; a desirable situation given the small number of bonds available by company.

Appendix A: MCMC Sampling Scheme

This Appendix presents a complete description of the Markov Chain Monte Carlo (MCMC) scheme used to sample from the posterior distribution of the proposed Bayesian population model. It also includes initial values and values for hyperparameters.

The proposed model without weights is given by:

$$P_{ib} \sim N\left(\Psi\left(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}\right), V_i\right)$$

$$\boldsymbol{\theta}_i \sim N\left(\boldsymbol{\mu}_i, \mathbf{S}\right) \qquad V_i \sim \text{Inv} - \chi^2(\nu, \sigma^2)$$

$$\boldsymbol{\mu}_i \sim G \equiv \sum_{h=1}^{\infty} w_h \delta\left(\boldsymbol{\mu}_h\right) \qquad \mathbf{S}^{-1} \sim Wishart(r, (r\mathbf{R})^{-1}) \qquad \sigma^2 \sim Ga(a_\tau, b_\tau)$$

$$G \sim DP(G_0, M)$$

$$M \sim Ga\left(a_m, b_m\right) \qquad G_0 \sim N\left(\mathbf{b}, \mathbf{B}\right)$$

$$\mathbf{b} \sim N(\mathbf{b}_0, \mathbf{B}_0) \qquad \mathbf{B}^{-1} \sim Wishart\left(w, (w\mathbf{W})^{-1}\right)$$

The posterior distribution does not have a closed form. A Markov chain Monte Carlo (MCMC) scheme to sample from the posterior distribution can be efficiently implemented since the kernel of the mixture and the base measure G_0 are both normally distributed (MacEachern and Müller 1998). Here we present a detailed description of such a scheme.

We start by introducing some notation and conventions. Let n be the number of term structures being estimated and g_i the number of bonds with term structure i. We need to keep track of which class a particular $\boldsymbol{\mu}_i$ belongs to since there is a positive probability that some of the mean vectors will equal each other. Let $\{\boldsymbol{\mu}_1^\star \ldots \boldsymbol{\mu}_k^\star\}$ be the set of distinct vectors from the set $\{\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n\}$. The induced groups on the model parameters $\boldsymbol{\theta}_i$ based on the value of the corresponding $\boldsymbol{\mu}_i$ are referred to as clusters. Thus, the number of clusters is k. We introduce an indicator vector $\mathbf{s} = (s_1, \ldots, s_n)$ mapping each $\boldsymbol{\theta}_i$ to a specific cluster, that is, $s_i = j$ iff $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j^\star$. The number of s_i for which $s_i = j$ is denoted as n_j. Finally, in this document the gamma distribution is parameterized in terms of the rate parameter, that is, the gamma density function is

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}\beta^\alpha exp(-\beta x)}{\Gamma(\alpha)};\ \alpha, \beta > 0,$$

with mean equal to α/β.

Using the notation introduced above, we describe the MCMC scheme for the model without weights.

1. Resampling s_i. We marginalize over $\boldsymbol{\mu}_i$. To write down the resultant distribution of s_i we need to introduce notation for the case when some $\boldsymbol{\mu}_i$ is removed from consideration. Such notation includes the following. $\mathbf{s}^{[i]} = (s_1, \ldots s_{i-1}, s_{i+1} \ldots, s_n)$. $\boldsymbol{\mu}^{[i]} = \{\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_{i-1}, \boldsymbol{\mu}_{i+1} \ldots, \boldsymbol{\mu}_n\}$. $n_j^{[i]}$ denotes the size of the jth cluster, it is equal to $n_{s_i} - 1$ for $j = s_i$ and n_j for other j. $k^{[i]}$ denotes the number of clusters, if removal of $\boldsymbol{\mu}_i$ shrinks the number of clusters so that

$k^{[i]} = k - 1$, then we relabel the kth cluster so that it becomes cluster s_i. This relabel is done by redefining $s_\ell = s_i$ for all ℓ with $s_\ell = k$, setting $\boldsymbol{\mu}^\star_{s_i} = \boldsymbol{\mu}^\star_k$, and removing $\boldsymbol{\mu}^\star_k$.

Using the notation introduced above, the distribution of s_i is multinomial with probability given by

$$p\left(s_i = j | \boldsymbol{\mu}^{[i]}, s^{[i]}, (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n), M, \mathbf{S}, \mathbf{b}, \mathbf{B}\right) \propto \begin{cases} n_j^{[i]} N(\boldsymbol{\theta}_i; \boldsymbol{\mu}^\star_j, \mathbf{S}) & j = 1, \cdots, k^{[i]} \\ M \int N(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \mathbf{S}) dG_0(\boldsymbol{\mu}) & j = k^{[i]} + 1. \end{cases}$$

See, for example, MacEachern and Müller (1998) for a derivation of this conditional distribution. The integral for the case $s_i = k^{[i]} + 1$ has a closed form:

$$\begin{aligned} \int N(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \mathbf{S}) dG_0(\boldsymbol{\mu}) &= \int N(\boldsymbol{\theta}_i; \boldsymbol{\mu}, \mathbf{S}) N(\boldsymbol{\mu}; \mathbf{b}, \mathbf{B}) d\boldsymbol{\mu} \\ &= N(\boldsymbol{\theta}_i; \mathbf{b}, \mathbf{S} + \mathbf{B}). \end{aligned}$$

If $s_i = k^{[i]} + 1$, we need to sample a new $\boldsymbol{\mu}^\star$ from the distribution

$$\begin{aligned} N(\boldsymbol{\mu}^\star; \alpha, \boldsymbol{\Sigma}) &\propto N\left(\boldsymbol{\theta}_i; \boldsymbol{\mu}^\star, \mathbf{S}\right) G_0(\boldsymbol{\mu}^\star) \\ &= N(\boldsymbol{\theta}_i; \boldsymbol{\mu}^\star_j, \mathbf{S}) N(\boldsymbol{\mu}^\star_j; \mathbf{b}, \mathbf{B}), \end{aligned}$$

where

$$\boldsymbol{\Sigma} = (\mathbf{S}^{-1} + \mathbf{B}^{-1})^{-1},$$

$$\boldsymbol{\alpha} = \boldsymbol{\Sigma}(\mathbf{S}^{-1}\boldsymbol{\theta}_i + \mathbf{B}^{-1}\mathbf{b}).$$

2. Resampling $\boldsymbol{\mu}_j^\star$.

$$\text{Prior} \qquad \boldsymbol{\mu}_j^\star \sim N(\mathbf{b}, \mathbf{B}),$$

$$\text{Likelihood} \quad L(\boldsymbol{\mu}_j^\star) = \prod_{i \in \Gamma_j} N\left(\boldsymbol{\theta}_i; \boldsymbol{\mu}_j^\star, \mathbf{S}\right),$$

$$\begin{aligned} \text{Posterior} \quad & \boldsymbol{\mu}_j^\star \sim N(\mathbf{b}^\star, \mathbf{B}^\star) \\ & \mathbf{B}^{\star -1} = \mathbf{B}^{-1} + n_j \mathbf{S}^{-1} \\ & \mathbf{b}^{\star -1} = \mathbf{B}^\star \left(\mathbf{B}^{-1}\mathbf{b} + \mathbf{S}^{-1}\left\{\sum_{i \in \Gamma_j} \boldsymbol{\theta}_i\right\}\right), \end{aligned}$$

where $\Gamma_j = \{i : s_i = j\}$.

3. Resampling $\boldsymbol{\theta}_i$:

$$\text{Prior} \qquad \boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}_i | \boldsymbol{\mu}_{s_i}^\star, \mathbf{S}),$$

$$\text{Likelihood} \quad L(\boldsymbol{\theta}_i) \propto \prod_{b=1}^{g_i} (V_i)^{-1/2} \exp\left(-\tfrac{1}{2V_i}\left\{P_{ib} - \Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})\right\}^2\right),$$

$$\text{Posterior} \quad P(\boldsymbol{\theta}_i | \cdots) \propto \text{Prior} \times \text{Likelihood},$$

The posterior of $\boldsymbol{\theta}_i$ does not have a closed form, thus we use a Metropolis-Hasting algorithm to update $\boldsymbol{\theta}_i$. Let $\boldsymbol{\theta}_i^t$ be the current point at the t iteration and $\boldsymbol{\theta}_i^{can}$ be the candidate value simulated from the proposal distribution $q(\boldsymbol{\theta}_i^{can} | \boldsymbol{\theta}_i^t)$. The acceptance probability of the candidate point is

$$min\left(1, \lambda\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^{can}\right)\right),$$

where

$$\lambda\left(\boldsymbol{\theta}_i^t, \boldsymbol{\theta}_i^{can}\right) = \frac{P(\boldsymbol{\theta}_i^{can}|\cdots)}{P(\boldsymbol{\theta}_i^t|\cdots)} \times \frac{q(\boldsymbol{\theta}_i^t|\boldsymbol{\theta}_i^{can})}{q(\boldsymbol{\theta}_i^{can}|\boldsymbol{\theta}_i^t)}.$$

The specification of a proposal distribution is often difficult since the posterior density is unknown. When the sample size is small, it is difficult to find good approximations to the posterior which could be used to set the parameters of the proposal distribution. To overcome this difficulty, we use the adaptive Metropolis (AM) algorithm introduced by Haario et al. (2001). The proposal distribution is a Gaussian distribution centered on the current state, whose covariance is calculated using all the previous states after a given burning period. Specifically, the proposal distribution is given by

$$q(\boldsymbol{\theta}_i^{can}|\boldsymbol{\theta}_i^t) \sim N(\boldsymbol{\theta}_i^t, \boldsymbol{\Sigma}_t),$$

with

$$\boldsymbol{\Sigma}_t = \begin{cases} \boldsymbol{\Sigma}_0 & t \leq t_0, \\ s_d \operatorname{cov}\left(\boldsymbol{\theta}_i^0, \ldots, \boldsymbol{\theta}_i^{t-1}\right) + s_d \epsilon \mathbf{I}_p & t > t_0, \end{cases}$$

where s_d and ϵ are positive constants, t_0 is a positive integer, and "cov" denotes the empirical covariance matrix. For this application, we set $s_d = 0.5$, $\epsilon = 0.00001$, and $t_0 = 2000$. The adaptation provided by the AM algorithm allows us to produce accurate estimators, even though we start with a rough approximation for the covariance matrix of the proposal distribution.

4. Resampling $\mathbf{S}$.

Prior $\quad \mathbf{S}^{-1} \sim Wishart(r, (r\mathbf{R})^{-1}),$

Likelihood
$$\begin{aligned} L(\mathbf{S}^{-1}) &\propto \textstyle\prod_{i=1}^{n} |\mathbf{S}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)\mathbf{S}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)^t\right] \\ &= |\mathbf{S}|^{-n/2} \exp\left[-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)\mathbf{S}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)^t\right], \end{aligned}$$

Posterior $\quad \mathbf{S}^{-1} \sim Wishart\left(r+n, [r\mathbf{R} + \textstyle\sum_{i=1}^{n}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)(\boldsymbol{\theta}_i - \boldsymbol{\mu}_i)^t]^{-1}\right),$

5. Resampling $\mathbf{b}$.

Prior $\quad \mathbf{b} \sim N(\mathbf{b}_0, \mathbf{B}_0),$

Likelihood $\quad L(\mathbf{b}) = \Pi_{j=1}^{k} N(\boldsymbol{\mu}_j^{\star}; \mathbf{b}, \mathbf{B}),$

Posterior
$$\begin{aligned} &\mathbf{b} \sim N(\mathbf{b}_1, \mathbf{B}_1) \\ &\mathbf{B}_1^{-1} = \mathbf{B}_0^{-1} + k\,\mathbf{B}^{-1} \\ &\mathbf{b}_1 = \mathbf{B}_1\left(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{B}^{-1}\left\{\textstyle\sum_{j=1}^{k}\boldsymbol{\mu}_j^{\star}\right\}\right), \end{aligned}$$

6. Resampling $\mathbf{B}$.

Prior $\quad \mathbf{B}^{-1} \sim Wishart\,(w, (w\mathbf{W})^{-1})\,,$

Likelihood $\quad L(\mathbf{B}^{-1}) = \prod_{j=1}^{k}(N(\boldsymbol{\mu}_j^{\star}; \mathbf{b}, \mathbf{B})$

$$\propto |\mathbf{B}|^{-k/2} \exp\left[-\tfrac{1}{2}(\textstyle\sum_{j=1}^{k}(\boldsymbol{\mu}_j^{\star} - \mathbf{b})\mathbf{B}^{-1}(\boldsymbol{\mu}_j^{\star} - \mathbf{b})^t\right],$$

Posterior $\quad \mathbf{B}^{-1} \sim Wishart\left(w + k, \left[w\mathbf{W} + \textstyle\sum_{j=1}^{k}(\boldsymbol{\mu}_j^{\star} - \mathbf{b})(\boldsymbol{\mu}_j^{\star} - \mathbf{b})^t\right]^{-1}\right).$

7. Resampling M. It is done by introducing a latent beta-distributed variable, η, as described in Escobar and West (1995). Let $p(\eta|M) = Beta(M+1, n)$ and $M \sim Ga(a_m, b_m)$, we have

$$p(M|\eta, k) = Ga(a'_m, b'_m),$$

with

$$\begin{aligned} b'_m &= b_m - log(\eta), \\ a'_m &= \begin{cases} a_m + k & \text{with probability } \pi = \frac{a_m+k-1}{a_m+k-1+nb'_m}, \\ a_m + k - 1 & \text{with probability } 1-\pi. \end{cases} \end{aligned}$$

8. Resampling V_i.

Prior $\quad V_i \sim \text{Inv} - \chi^2(\nu, \sigma^2),$

Likelihood $\quad L(V_i) = \prod_{b=1}^{g_i} N\left(P_{ib}|\Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}), V_i\right),$

Posterior $\quad V_i \sim \text{Inv} - \chi^2(\nu + 1, \frac{1}{\nu+1}\left\{\nu\sigma^2 + \sum_{b=1}^{g_i}\{P_{ib} - \Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})\}^2\right\}).$

9. Resampling σ^2.

$$\text{Prior} \qquad \sigma^2 \sim Ga(a_\tau, b_\tau),$$

$$\text{Likelihood} \quad L(\sigma^2) = \prod_{i=1}^{n} \text{Inv} - \chi^2(V_i|\nu, \sigma^2),$$

$$\text{Posterior} \quad \sigma^2 \sim Ga\left(a_\tau + (n\nu)/2, b_\tau + (\nu/2)\sum_{i=1}^{n} V_i^{-1}\right).$$

Weights, ω_{ib}, are introduced into our model by modifying the scale parameter in the likelihood as follows:

$$P_{ib} \sim N\left(\Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}), V_i(\omega_{ib})^{-1}\right).$$

Which is equivalent to

$$\sqrt{\omega_{ib}} P_{ib} \sim N\left(\sqrt{\omega_{ib}}\, \Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}), V_i\right).$$

Hence, the only change needed in the MCMC scheme described above when weights are considered is to multiply P_{ib} and $\Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})$ by $\sqrt{\omega_{ib}}$ in steps 3 and 8.

Finally, we describe the procedure we followed in order to set hyperparameters and initial values. We assume that we can estimate the mean and covariance matrix of the population of parameters θ_i. In this document we have used the following vector to estimate the population mean

$$\breve{\boldsymbol{\mu}} = (-143.93, -283.08, 56.02, 93.75)^T,$$

while for the covariance matrix we used

$$\check{\boldsymbol{S}} = \begin{pmatrix} 582.93 & 1019.38 & -638.38 & -762.05 \\ 1019.38 & 2341.63 & -1524.80 & -1237.71 \\ -638.38 & -1524.80 & 1001.85 & 749.14 \\ -762.05 & -1237.71 & 749.14 & 1073.54 \end{pmatrix}.$$

Such estimators were obtained by computing the sample mean and covariance matrix, respectively, of the estimated parameters θ describing the term structure of 697 U.S. companies. The vector of parameters θ were estimated by fitting the proposed Bayesian population model to a bond data set available through the Yahoo's bond screener (`http://screen.yahoo.com/bonds.html`). Such data set includes bond data for 1877 U.S. corporate bonds with information corresponding to May 29, 2009.

Based on $\check{\boldsymbol{\mu}}$ and $\check{\boldsymbol{S}}$ as well as the posterior results obtained using the Yahoo data set, the hyperparameters and initial values are set as follows.

Hyperparameters

$b_0 = \check{\boldsymbol{\mu}}$, $B_0 = \check{\boldsymbol{S}}$, $w = 15$, $W = \check{\boldsymbol{S}}/20$, $r = 10$, $R = \check{\boldsymbol{S}}/10$, $a_\tau = 0.5$, $b_\tau = 1$, $a_m = 1$, and $b_m = 1$.

Initial Values.

$b = \check{\boldsymbol{\mu}}$, $B = \check{\boldsymbol{S}}$, $M = 1.0$, $S = \check{\boldsymbol{S}}$, $\sigma^2 = 0.5$, $\mu_i = \check{\boldsymbol{\mu}}$, $s_i = i$, $V_i = (\nu/(\nu-2)) * \sigma^2$, and for the Metropolis-Hasting algorithm to update θ_i we set $\Sigma_0 = (1/4) * \check{\boldsymbol{S}}$.

Appendix B: Dependence on Covariates, MCMC and Predictive Density

This appendix includes a detailed explanation of the MCMC scheme to sample from the posterior distribution of the model described in Section 2.5.1. It also explains how to use a sample from the posterior distribution to approximate the posterior density.

We consider the following definitions and conventions. Let p be the number of parameters in the model, $\alpha^\star = \{\alpha_j^\star\}$ be the set of unique matrices from the set $\{\alpha_i\}$, k be the size of the set $\alpha^\star$, n the number of patients, g_i the number of concentration-time observations available for the ith patient, and $\alpha^{[i]} = \{\alpha_1, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_n\}$. As described in Section 2.5.1, when the model includes categorical covariates the base measure has the form

$$G_0(\alpha_j^\star) = N(b, B) \times \Pi_{k=2}^{c(1)} N(0, B_1) \times \cdots \times \Pi_{k=2}^{c(d)} N(0, B_d).$$

which can be rewritten as

$$G_0(\alpha_j^\star) = N(b, B) \times N(0, \tilde{B}_1) \times \cdots \times N(0, \tilde{B}_d),$$

where

$$\tilde{B}_\ell = \begin{pmatrix} B_\ell & 0 & \cdots & 0 & 0 \\ 0 & B_\ell & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & B_\ell & 0 \\ 0 & 0 & \cdots & 0 & B_\ell \end{pmatrix},$$

with B_ℓ appearing $c(\ell) - 1$ times in $\tilde{B}_\ell$, for $\ell = 1 \ldots d$.

Let $\tilde{\alpha}_j^\star$ be the column vector obtained by writing each column of $\alpha_j^\star$ one after the other. An equivalent way to write the base measure in terms of $\tilde{\alpha}_j^\star$ is

$$G_0(\tilde{\alpha}_j^\star) \sim N\left(\tilde{b} = \begin{pmatrix} b \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \tilde{B} = \begin{pmatrix} B & 0 & \cdots & 0 & 0 \\ 0 & \tilde{B}_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \tilde{B}_{d-1} & 0 \\ 0 & 0 & \cdots & 0 & \tilde{B}_d \end{pmatrix}\right).$$

Set

$$F_i = d_i^t \otimes I_p,$$

where I_p is the $p \times p$ identity matrix, and d_i is the design vector for the ith subject. Introducing F_i is convenient because of the identity

$$\alpha_j^\star d_j = F_i \tilde{\alpha}_j^\star.$$

Using the notation introduced above, the corresponding MCMC scheme is as follows.

1. Resampling s_i. The posterior distribution of s_i is given by

$$Pr(s_i = j|\ldots) \propto \begin{cases} \frac{n_j^{[i]}}{M+n-1} N(\theta_i; \alpha_j^\star d_i, S) & j = 1, \cdots, k^{[i]}, \\ \\ \frac{M}{M+n-1} \int N(\theta_i; \alpha d_i, S) dG_0(\alpha) & j = k^{[i]} + 1. \end{cases}$$

The integral $\int N(\theta_i; \alpha d_i, S) dG_0(\alpha)$ has the following closed form:

$$\begin{aligned} \int N(\theta_i; \alpha d_i, S) dG_0(\alpha) &= \int N(\theta_i; F_i\tilde{\alpha}, S) dG_0(\tilde{\alpha}) \\ &= \int N(\theta_i; F_i\tilde{\alpha}, S) N(\tilde{\alpha}; \tilde{b}, \tilde{B}) d\tilde{\alpha} \\ &= N(\theta_i; F_i\tilde{b}, S + F_i\tilde{B}F_i^T). \end{aligned}$$

If $s_i = k^{[i]} + 1$, we need to sample a new $\alpha_j^\star$ from the distribution

$$N(\alpha; \mu, \Sigma) \propto N(\theta_i; F_i\tilde{\alpha}_j^\star, S) \times N(\tilde{\alpha}_j^\star; \tilde{b}, \tilde{B}),$$

where

$$\begin{aligned} \Sigma &= (F_i^t S^{-1} F_i + \tilde{B}^{-1})^{-1}, \\ \mu &= \Sigma(F_i^t S^{-1} \theta_i + \tilde{B}^{-1}\tilde{b}). \end{aligned}$$

2. Resampling $\alpha_j^\star$. The posterior distribution of $\alpha_j^\star$ is given by

$$p(\alpha_j^\star | s, (\theta_1, \cdots, \theta_n), b, B, S, \rho) \propto \prod_{i \in \Gamma_j} N(\theta_i; \alpha_j^\star d_i, S) G_0(\alpha_j^\star),$$

where $\Gamma_j = \{i : s_i = j\}$, and

$$G_0(\alpha_j^\star) = N(b, B) \times N(0, \tilde{B}_1) \times \cdots \times N(0, \tilde{B}_d).$$

We can update $\tilde{\alpha}_j^\star$ in the linear regression setting by noting that

$$\theta_i = F_i\tilde{\alpha}_j^\star + \omega_i, \quad \omega_i \sim N(0, S) \text{ for } i \in \Gamma_j,$$

$$\tilde{\alpha}_j^\star \sim N(\tilde{b}, \tilde{B}),$$

therefore, $\tilde{\alpha}_j^\star$ can be updated sequentially using standard results of Normal theory as follows.

Let $\tilde{\theta}_j$ be the vector of θ_i's such that $i \in \Gamma_j$, that is, $\tilde{\theta}_j = \left(\theta_{j1}, \theta_{j2}, \ldots, \theta_{jn_j}\right)^t$ where $j1, \cdots, jn_j$ are the ordered indexes in Γ_j. Similarly, we define $\omega_j = \left(\omega_{j1}, \ldots, \omega_{jn_j}\right)^t$ and

$$\tilde{F}_j = \begin{pmatrix} d_{j1}^t \\ \vdots \\ d_{jn_j}^t \end{pmatrix} \otimes I_p.$$

Rewriting the linear regression model, we have

$$\tilde{\theta}_j = \tilde{F}_j\tilde{\alpha}_j^\star + \tilde{\omega}_j, \quad \tilde{\omega}_j \sim N(0, I_{nj} \otimes S),$$

$$\tilde{\alpha}_j^\star \sim N(\tilde{b}, \tilde{B}).$$

The posterior distribution of $\tilde{\alpha}_j^\star$ is $N(Cc, C)$ where

$$C^{-1} = \tilde{F}_j^T(I_{nj} \otimes S)^{-1}\tilde{F}_j + \tilde{B}^{-1} = \tilde{F}_j^T(I_{nj} \otimes S^{-1})\tilde{F}_j + \tilde{B}^{-1}$$

and

$$c = \tilde{F}_j^T(I_{nj} \otimes S^{-1})\tilde{\theta}_j + \tilde{B}^{-1}\tilde{b}$$

3. Resampling θ_i:

Prior $\quad \theta_i \sim N(\theta_i|\alpha^{\star}_{s_i} d_i, S),$

Likelihood $\quad L(\theta_i) \propto \prod_{b=1}^{g_i} (V_i)^{-1/2} \exp\left(-\frac{1}{2V_i}\{P_{ib} - \Psi(\theta_i, CF_{ib})\}^2\right),$

Posterior $\quad P(\theta_i|\cdots) \propto$ Prior $\times$ Likelihood,

Since the posterior of θ_i does not have a closed form, we use a Metropolis-Hasting algorithm to update θ_i. We use the proposal distribution corresponding to the adaptive Metropolis (AM) algorithm (Haario et al. (2001)). The details are similar to those described in Appendix A.

4. Resampling S.

Prior $\quad S^{-1} \sim Wishart(r, (rR)^{-1}),$

Likelihood
$$\begin{aligned} L(S^{-1}) &\propto \Pi_{i=1}^{n} |S|^{-1/2} \exp[-\tfrac{1}{2}(\theta_i - F_i\tilde{\alpha}_i)S^{-1}(\theta_i - F_i\tilde{\alpha}_i)^t] \\ &= |S|^{-n/2} \exp[-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(\theta_i - F_i\tilde{\alpha}_i)S^{-1}(\theta_i - F_i\tilde{\alpha}_i)^t], \end{aligned}$$

Posterior $\quad S^{-1} \sim Wishart(r+n, [rR + \sum_{i=1}^{n}(\theta_i - F_i\tilde{\alpha}_i)(\theta_i - F_i\tilde{\alpha}_i)^t]^{-1}),$

5. Resampling b.

$$\text{Prior} \qquad b \sim N(b_0, B_0),$$

$$\text{Likelihood} \quad L(b) = \Pi_{j=1}^{k} N(\alpha_j^\star[,1]; b, B),$$

$$\text{Posterior} \quad b \sim N(b_1, B_1)$$

$$B_1^{-1} = B_0^{-1} + kB^{-1}$$

$$b_1 = B_1(B_0^{-1}b_0 + kB^{-1}\bar{\alpha}^\star[,1]),$$

where $\alpha_j^\star[,1]$ stands for the first column of the matrix $\alpha_j^\star$, and $\bar{\alpha}^\star[,1] = (1/k)\sum_{j=1}^{k} \alpha_j^\star[,1]$.

6. Resampling B.

$$\text{Prior} \qquad B^{-1} \sim Wishart(w, (wW)^{-1}),$$

$$\text{Likelihood} \quad L(B^{-1}) = \Pi_{j=1}^{k}(N(\alpha_j^\star[,1]; b, B)$$

$$\propto |B|^{-3k/2} \exp\left[-\tfrac{1}{2}(\textstyle\sum_{j=1}^{k}(\alpha_j^\star[,1] - b)B^{-1}(\alpha_j^\star[,1] - b)^t\right],$$

$$\text{Posterior} \quad B^{-1} \sim Wishart(w + k, \left[wW + \textstyle\sum_{j=1}^{k}(\alpha_j^\star[,1] - b)(\alpha_j^\star[,1] - b)^t\right]^{-1}).$$

7. Resampling B_ℓ. Let $j_2, j_3, \ldots, j_{c(\ell)}$ be the indices of the columns in $\alpha_j^\star$ corres-

ponding to the offsets of the ℓth categorical covariate.

Prior $\quad B_\ell^{-1} \sim Wishart(w, (wW)^{-1}),$

Likelihood $\quad L(B_\ell^{-1}) = \Pi_{j=1}^{k}(N(\alpha_j^\star[,j_2]; 0, B_\ell^{-1}) \times \cdots \times N(\alpha_j^\star[,j_{c(\ell)}]; 0, B_\ell^{-1}))$

$$\propto |B_\ell|^{-(c(\ell)-1)k/2} \exp[-\tfrac{1}{2}(\textstyle\sum_{j=1}^{k}(\alpha_j^\star[,j_2] - 0)B_\ell^{-1}(\alpha_j^\star[,j_2] - 0)^t +$$

$$\cdots + (\alpha_j^\star[,j_{c(\ell)}] - 0)B_\ell^{-1}(\alpha_j^\star[,j_{c(\ell)}] - 0)^t))],$$

Posterior $\quad B_\ell^{-1} \sim Wishart(w + (c(\ell) - 1)k, [wW + \textstyle\sum_{j=1}^{k}(\alpha_j^\star[,j_2])(\alpha_j^\star[,j_2])^t +$

$$\cdots + (\alpha_j^\star[,j_{c(\ell)}])(\alpha_j^\star[,j_{c(\ell)}])^t]^{-1}).$$

8. Resampling M , V_i and σ^2 is done as described in Appendix A.

Finally, we explain how to estimate the corresponding predictive distribution. If we denote the data used to fit the model as Y, the new vector of covariates as x_{n+1}, and the parameters in the term structure model as Θ, then the predictive density is given by

$$\begin{aligned} p(\theta_{n+1}|x_{n+1}, Y) &= E_{\Theta|Y}\left[p(\theta_{n+1}|x_{n+1}, Y, \Theta)\right] \\ &\approx \frac{1}{T}\sum_{l} p\left(\theta_{n+1}|x_{n+1}, Y, \Theta^{(t)}\right), \end{aligned}$$

where $\left\{\Theta^{(t)}|t = 1, \ldots, N\right\}$ is a sample of size N from the posterior distribution and

$$\begin{aligned} p\left(\theta_{n+1}|x_{n+1}, Y, \Theta^{(t)}\right) \propto & \sum_{j=1}^{k^{(t)}} n_j^{(t)} N\left(\theta_{n+1}; \alpha_j^{\star(t)} d_{n+1}, S^{(t)}\right) \\ & + M^{(t)} \int N(\theta_{n+1}; \alpha\, d_{n+1}, S^{(t)}) dG_0^{(t)}(\alpha). \end{aligned} \tag{4.1}$$

In order to sampling from the predictive distribution, we can draw a sample from the posterior distribution, and use it to sample from the mixture $p\left(\theta_{n+1}|x_{n+1}, Y, \Theta^{(t)}\right)$.

Appendix C: Normally Distributed Prices

The term structure estimation model introduced in this document assumes that the bond prices follow a t-distribution (see equation (2.7)). Hence, it allows us to study how the performance of the model varies when the distribution of the prices has heavy or light tails, respectively. In a less flexible version of our term structure model, we assumed normally distributed prices. In general, similar in-sample price residuals are obtained when using either normally or t distributed prices (see Table 4.1).

Method	Minimum	Q1	Median	Mean	Q3	Maximum
t(3)	0.00	0.23	0.62	1.14	1.58	18.16
t(10)	0.00	0.25	0.59	1.14	1.52	18.18
Normal	0.00	0.27	0.63	1.23	1.53	22.69

Table 4.1 : Summary statistics for absolute price residuals under different distributional assumptions for the prices. Q1 and Q3 denote the first and third quartiles, respectively. The estimators of the term structure of individual firms computed with our Bayesian population model have similar in-sample performance when the prices are assumed to be normally distributed (Normal) or when they follow a t distribution with degrees of freedom 3 (t(3)) or 10 (t(10)), respectively.

When considering out-of-sample test, however, the distributional assumption on the prices does lead to different results. Specifically, lighter tails in the distribution

of the prices seems to provide a better fit. The out-of-sample tests are based on the partitions described in Section 2.3 (see Table 4.2 in this appendix).

	RMSPE			MAPE		
	(2)	(3)	(4)	(2)	(3)	(4)
t(3)	4.17	6.41	2.69	2.71	3.83	1.64
t(10)	3.30	3.81	1.92	2.50	2.65	1.33
Normal	3.28	3.61	1.74	2.44	2.56	1.25

Table 4.2 : Out-of-sample statistics under different distributional assumptions for the prices. RMSPE stands for root mean squared prediction error while MAPE stands for mean absolute prediction error. The table shows the average of the RMSPE and MAPE over partitions with (m, k) having the same m (number in parenthesis). See Section 2.3 for a description of the partitions. In all cases the estimators of the term structures of individual firms produced with our Bayesian population model (BPM) outperforms those obtained with the single-curve approach (Single).

Regarding the MCMC sampling scheme under normally distributed prices, the algorithm is very similar to the one described in Appendix A. We just need to replace V_i and σ^2 for a precision parameter τ which correspond to the likelihood

$$P_{ib} \sim N\left(\Psi\left(\boldsymbol{\theta}_i, \mathbf{CF}_{ib}\right), \tau^{-1}\right),$$

and use the following resampling step:

* Resampling τ.

Prior $\quad \tau \sim Ga(a_\tau, b_\tau),$

Likelihood $\quad L(\tau) \propto \prod_{i=1}^{n} \prod_{b=1}^{g_i} (\tau)^{1/2} \exp\left(-\frac{1}{2}\tau \left\{P_{ib} - \Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})\right\}^2\right),$

Posterior $\quad \tau \sim Ga\left(a_\tau + \frac{1}{2}\sum_{i=1}^{n} g_i\ ,\right.$

$$\left. b_\tau + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{g_i} \left\{P_{ib} - \Psi(\boldsymbol{\theta}_i, \mathbf{CF}_{ib})\right\}^2\right).$$

Appendix D: Software Implementation

This appendix briefly describes the software implementation of the proposed term structure estimation model described in Appendix A.

The routines are written in C. We used the functions in the GNU Scientific Library (GSL) for defining and manipulating vectors and matrices as well as for generating random numbers. An advantage of using the numerical library GSL is that it has a detailed documentation freely available at `http://www.gnu.org/software/gsl/`. Having such a documentation is advantageous because it makes it easier for new users to understand and to implement our term structure estimation method.

The C code is a straightforward implementation of the MCMC algorithm described in Appendix A. Specifically, it includes a function *main* that calls other auxiliary functions corresponding to each resampling step. There are as many auxiliary functions as resampling steps in the MCMC algorithm. Such an organization of the code facilitates its readability.

For additional information regarding the code, please contact the Department of Statistics at Rice University.

Bibliography

Bliss, R. (1997), "Testing Term Structure Estimation Methods," *Advances in Futures and Options Research*, 9, 197–232.

Burr, D. and Doss, H. (2005), "A Bayesian Semiparametric Model for Random-Effects Meta-Analysis," *Journal of the American Statistical Association*, 100, 242–251.

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2006), "Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures," *IEEE Transactions on Signal Processing*, 56, 71–84.

Carota, C. and Parmigiani, G. (2002), "Semiparametric Regression for Count Data," *Biometrika*, 89, 265–285.

Cifarelli, D. and Regazzini, E. (1978), "Nonparametric Statistical Problems Under Partial Exchangeability. The use of Associative Means," *Annali del Instituto di Matematica Finianziara dell Universita di Torino*, 12, 1–36.

De Iorio, M., Johnson, W., Muller, P., and Rosner, G. (2009), "Bayesian Nonparametric Nonproportional Hazards Survival Modeling," *Biometrics*, 65, 762–771.

De Iorio, M., Müller, P., Rosner, G., and MacEachern, S. (2004), "An ANOVA Model

for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215.

Duffee, G. (1996), "On Measuring Credit Risks of Derivative Instruments," *Journal of Banking and Finance*, 20, 805–833.

Duffie, D. and Singleton, K. (1999), "Modeling Term Structures of Defaultable Bonds," *Review of Financial Studies*, 12, 687–720.

Dunson, D. (2009), "Bayesian Nonparametric Hierarchical Modeling," *Biometrical Journal*, 51, 273–284.

Dunson, D. and Park, J. (2008), "Kernel Stick-Breaking Processes," *Biometrika*, 95, 307–323.

Dunson, D., Pillai, N., and Park, J. (2007), "Bayesian Density Regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 163–183.

Elton, E., Gruber, M., Agrawal, D., and Mann, C. (2004), "Factors Affecting the Valuation of Corporate Bonds," *Journal of Banking & Finance*, 28, 2747 – 2767.

Escobar, M. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230.

Ferstl, R. and Hayden, J. (2008), "Zero Coupon Yield Curve Estimation with the Package termstrc," preprint. Available at SSRN: http://ssrn.com/abstract=1307149.

Fuentes-García, R., Mena, R., and Walker, S. (2009), "A Nonparametric Dependent Process for Bayesian Regression," *Statistics and Probability Letters*, 79, 1112–1119.

Gelfand, A., Kottas, A., and MacEachern, S. (2005), "Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing." *Journal of the American Statistical Association*, 100, 1021–1036.

Gelman, A., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, CRC press, 2nd ed.

Griffin, J. and Steel, M. (2004), "Semiparametric Bayesian Inference for Stochastic Frontier Models," *Journal of Econometrics*, 123, 121–152.

— (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An Adaptive Metropolis Algorithm," *Bernoulli*, 7, 223–242.

Houweling, P., Hoek, J., and Kleibergen, F. (2001), "The Joint Estimation of Term Structures and Credit Spreads," *Journal of Empirical Finance*, 8, 297–323.

Hull, J. and White, A. (1995), "The Impact of Default Risk on the Prices of Options and other Derivative Securities," *Journal of banking and finance*, 19, 299–322.

Ioannides, M. (2003), "A Comparison of Yield curve estimation techniques using UK data," *Journal of Banking and Finance*, 27, 1–26.

Jarrow, R., Ruppert, D., and Yu, Y. (2004), "Estimating the Interest Rate Term Structure of Corporate Debt With a Semiparametric Penalized Spline Model," *Journal of the American Statistical Association*, 99, 57–66.

Jarrow, R. and Turnbull, S. (1995), "Pricing Derivatives on Financial Securities Subject to Credit Risk," *The Journal of Finance*, 50, 53–85.

Kim, S., Tadesse, M., and Vannucci, M. (2006), "Variable Selection in Clustering via Dirichlet Process Mixture Models," *Biometrika*, 93, 877–893.

Li, M. and Yu, Y. (2005), "Estimating the Interest Rate Term Structures of Treasury and Corporate Debt with Bayesian Penalized Splines," *Journal of Data Science*, 3, 223–240.

MacEachern, S. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

MacEachern, S. and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.